Automation of Summarization Evaluation Methods and their Application to the Summarization Process

Thade Nahnsen

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2011

Abstract

Summarization is the process of creating a more compact textual representation of a document or a collection of documents. In view of the vast increase in electronically available information sources in the last decade, filters such as automatically generated summaries are becoming ever more important to facilitate the efficient acquisition and use of required information. Different methods using natural language processing (NLP) techniques are being used to this end. One of the shallowest approaches is the clustering of available documents and the representation of the resulting clusters by one of the documents; an example of this approach is the Google News website. It is also possible to augment the clustering of documents with a summarization process, which would result in a more balanced representation of the information in the cluster, NewsBlaster being an example. However, while some systems are already available on the web, summarization is still considered a difficult problem in the NLP community. One of the major problems hampering the development of proficient summarization systems is the evaluation of the (true) quality of system-generated summaries. This is exemplified by the fact that the current state-of-the-art evaluation method to assess the information content of summaries, the Pyramid evaluation scheme, is a manual procedure.

In this light, this thesis has three main objectives.

1. The development of a fully automated evaluation method. The proposed scheme is rooted in the ideas underlying the Pyramid evaluation scheme and makes use of deep syntactic information and lexical semantics. Its performance improves notably on previous automated evaluation methods.

2. The development of an automatic summarization system which draws on the conceptual idea of the Pyramid evaluation scheme and the techniques developed for the proposed evaluation system. The approach features the algorithm for determining the pyramid and bases importance on the number of occurrences of the variable-sized contributors of the pyramid as opposed to word-based methods exploited elsewhere.

3. The development of a text coherence component that can be used for obtaining the best ordering of the sentences in a summary.

Acknowledgements

I would most sincerely like to thank my supervisor Claire Grover for her unflagging guidance, support, and encouragement throughout the course of completing this thesis. I would also like to convey my appreciation to my second supervisor Mirella Lapata for her valuable input and thank my DDD committee members Ewan Klein and Jean Carletta for their constructive comments and advice. I am likewise very grateful to my examiners Steve Renals and Massimo Poesio for making my viva an interesting and truly refreshing experience.

My studies at the University of Edinburgh, and thus this thesis, could not have taken the shape the have done without generous funding by the EPSRC, School of Informatics, and ICCS – thank you for granting me this remarkable experience.

On a personal note, I would like to thank my girlfriend Sascha for her help and support during the many discussions about and revisions of my thesis. Thank you for being my "rock!"

Last but by no means least, I would like to thank my family for their backing and patience during my various academic pursuits in the last decade.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Thade Nahnsen*)

To my parents

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this age of digitalization, one has access to enormous amounts of information at the mere click of a button. Routinely, however, much of that information is redundant because content is (frequently) repeated many times over, or one has preliminary knowledge on a query and requires specifics on a particular detail rather than a broad overview of every piece of information conceivably available given some search input. This brief preamble already hints at many of the intricacies and complexities associated with this subject area. For instance, how precise is the individual's input or search query in general? What does (s)he already know ahead of making a particular query? How much is too much? And so forth.

Many different approaches have been used and are continually being put forward to obtain or, "pull out," the information that is actually sought. The most prominent approaches among them are information retrieval, information extraction, and automatic summarization. Each works on a different level of granularity, or coarseness. The crudest of the approaches is information retrieval, the aim of which is the (rudimentary) recovery of documents that are most relevant to a particular set of keywords. Information extraction and automatic summarization are more sophisticated in that they strive to retrieve the appropriate (specific) parts of documents for a given query.

In the context of a "coarseness hierarchy," the approach following information retrieval would be automatic summarization. Its objective is to extract the most pertinent information in a document or a collection of documents and return them in the form of coherent natural language text. Paralleling information retrieval, the final product is a document. Yet, rather than an essentially random set of documents related to the search query, the result is a single file containing a summary of all relevant information, say, from the set of documents obtained via the retrieval method. At the top of the hierarchy, one finds information extraction. As the term implies, the information obtained is a specific nugget given some precise information requirement. Someone working on a threat assessment regarding imminent terrorist activity, for example, may need to find out "the (exact) number of people killed in terrorist attack X on date Y in location Z."

Plainly, the input requirement(s) and ultimate output for each of these approaches – and most others available – are quite distinct. In the case of summarization in particular, clear-cut indications of the information to be extracted tend not to be available, entailing a correspondingly vague output. What is more, in order to be able to make efficient use of a summary, it has to be coherent and structured in a way accessible to the individual making the query, with the most important information clearly and

succinctly presented. The purpose of the work described in this thesis, broadly speaking, is the assessment of the quality of summaries automatically generated by existing methods and, given this valuation, the enhancement of these methods.

The field of automatic summarization is an exceedingly varied field of research. As in many other fields in informatics, one of the main foci is the evaluation of automatically generated summaries. The remainder of this chapter is devoted to detailing the shortcomings of contemporary approaches and outlining the steps employed to develop them in a number of useful directions.

In notable contrast to the field's underlying premise, the standard approach to evaluating the informational content of automatically generated (also called peer) summaries – the Pyramid evaluation scheme (Nenkova and Passanneau, 2004) – has to be run *manually*. While its accuracy and integrity are generally accepted to be close to ideal, it would undoubtedly be preferable, not to say more efficient, to have an automatic system of evaluation based (only) on a set of human reference summaries, such as is, for instance, done by ROUGE (Lin and Hovy, 2003). Thus, the first part of this thesis investigates the feasibility of fully automating the methodology underlying the Pyramid evaluation method, the objective being a higher correlation of the ensuing scores to the manual Pyramid score than those achieved by current automatic evaluation methods, ROUGE in particular. To this end, in a first step, given a manually constructed pyramid, a system is constructed to automatically evaluate a peer summary against the existing pyramid. Based on this partially automated evaluation, in a second step, a clustering approach combines sub-sentential units; the clusters, in turn, are combined on the basis of proximity constraints. The result of this grouping process is an automatically created pyramid. Owing to its emphasis on informational content as opposed to surface realizations, the ensuing evaluation scheme indeed correlates better with the manual Pyramid evaluation than ROUGE.

Building on this endeavor, the second part of the thesis explores the usefulness of the concepts underlying the evaluation scheme for the purpose of *generating* summaries. Note that, in this regard, most multi-document summarization systems and evaluation schemes are based on the notion that the most repeated information is most important. In summarization systems, this is usually translated into scores for the importance of either sentences or other clause-sized units. Problematically, however, the appropriate unit of summarization is usually not obvious. It may be based on whole sentences, discourse units, word triples, as well as n-grams, though in all cases, the unit is fixed. The system proposed in this part of the thesis tackles this issue by ex-

ploiting a variation of one of the components developed for the evaluation scheme to determine the most important units irrespective of their individual size (i.e., the units may vary in size). To be precise, I utilize the component of the evaluation system responsible for determining the pyramid units to select units from the original documents and gauge their relative importance based on the number of occurrences in the original documents as well as a (small) number of other characteristics of the ensuing variable-sized informational units. This would make it possible to select surface realizations that minimize the number of words required to capture the information, thereby enabling the system to represent more information in the summary. The resulting system compares favorably to other automatic summarization systems.

The remaining challenge confronted in this thesis relates to the readability, coherence, and structure of the summaries created via the proposed scheme. Having constructed a summarization method for determining the most important information in a collection of documents, it is crucial to determine the optimal presentation of the relevant information. To this end, the final part of this thesis seeks to establish which surface realizations for particular pieces of information should be used, and in which order the information should be presented. As before, the approach to this objective is two-tiered. First, a method is devised to determine the quality of an ordering of sentences relative to an alternative ordering of the same sentences. Second, using the resulting information, an optimal ordering of a set of sentences in constructed.

In sum, the objective of the work presented in this thesis is the introduction of a streamlined "solution" to the main difficulties faced in the generation and evaluation of automatic summaries: informational content, coherence, and structure. In their essence, the proposals constitute an automated version of the widely used Pyramid evaluation scheme, and include a novel approach to sentence ordering. From a technical point of view, the contributions of this thesis are as follows:

- an automation of the Pyramid evaluation scheme based on syntactic and semantic analysis;
- a method for sentence ordering based on shallow syntactic information;
- an algorithm and data structure to determine complex dependencies between different surface realizations of similar information; and
- an algorithm for determining the quality of the ordering of the sentences in a summary.

The thesis is structured as follows. Chapter 2 provides a general overview of the areas of automatic summarization related to the work presented in this thesis. On this foundation, Chapters 3 and 4 present the two-tiered development of the automatic method for evaluating summarization systems. Drawing on the discussion of the evaluation scheme, Chapter 5 introduces a generalized version of the scheme to be used to generate summaries. As a final step, Chapter 6 outlines an algorithm to determine the optimal ordering of sentences in multi-document summarization systems. Chapter 7 brings the thesis to a close with ideas for future work and concluding remarks.

Chapter 2

Related Work

To place the work presented in this thesis into its context within the field of automatic summarization, this chapter features a survey of the current state-of-the-art in summarization research, paying particular attention to the research motivating and informing the present work. It starts out with a characterization of the task of automatic summarization (Section 2.1), which is followed by a classification of the summarization task and the required processing steps along a number of dimensions relating to the nature of their content, depth, and complexity (Section 2.2). These fairly general introductory sections lay the foundation for a overview of the status quo of automatic summarization systems (Section 2.3). The final constituents of this chapter are three sections reviewing work explicitly related to the methodology developed in this thesis. The systems are grouped according to their main contributions: (a) the selection and extraction of relevant information (Section 2.4), (b) sentence ordering (Section 2.5), and (c) the evaluation of summarization systems (Section 2.6).

2.1 What is Automatic Summarization?

Etymologically, the term "summarization" has its roots in the Latin word "summa," which designates a concise recitation of the salient facts of some event or acquired information. The suffix "-ation," from "-ātiō," moreover, indicates a process or action. Correspondingly, summarization denotes the often complicated process of "restating the essence of text or an experience in as few words as possible or in a new, yet efficient, manner" (Wormeli, 2005). Note that this definition is not restricted to the laborious task recurrently imposed on students in English class, nor is it limited to textual information. Rather, it also encompasses non-manual approaches, and can apply to a wide range of circumstances.

As the term implies, (the field of) automatic summarization is concerned with automating the summarization task. For the most part, relevant work tends to focus on summarizing textual information, though the field also comprises areas dedicated to such varied undertakings as video (e.g., Ma et al., 2002) and meeting summarization (e.g., Murray et al., 2005). While the fundamental idea is similar in all cases, the various research areas clearly pose different challenges and may require the incorporation of tools from speech recognition or the integration of several different methodologies to achieve their ends (such as combining document and video summarization techniques to be able to summarize an entire meeting).

The present work focuses on the summarization of textual information dispersed

across a multitude of documents, or multi-document automatic summarization. The basic task is illustrated in Figure 2.1. Given an "input" of multiple documents on the same (general) topic, as identified by the newswire articles depicted on the left-hand side, the objective is the generation of a single document containing only the most relevant information from the original set of sources.

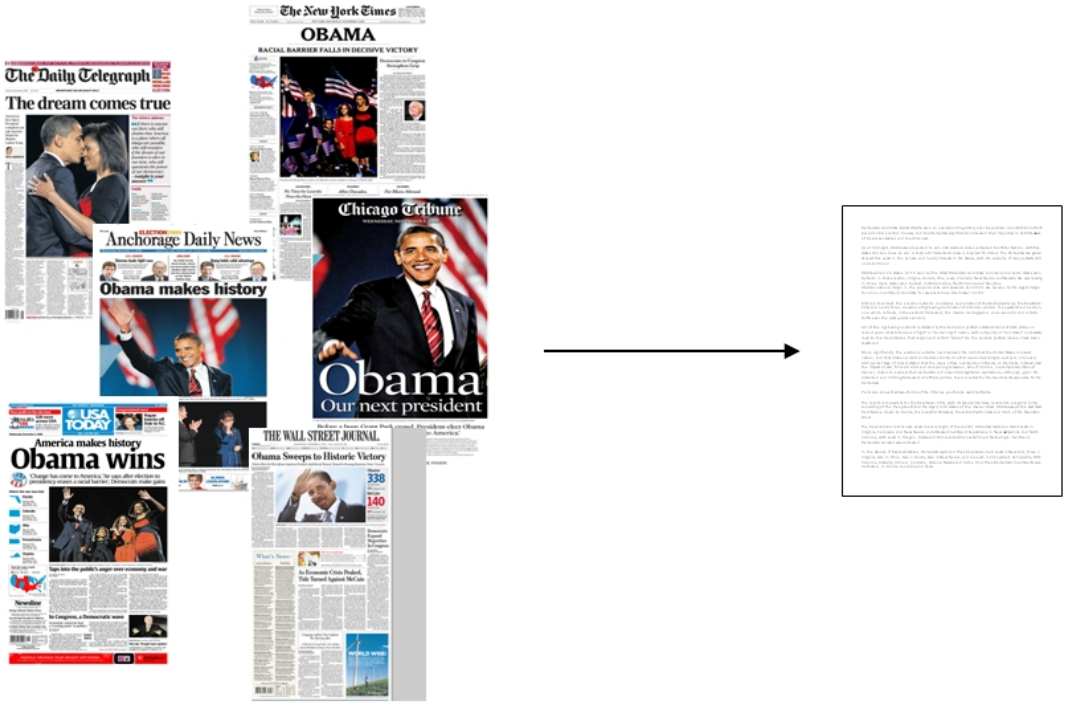

Figure 2.1: The task of multi-document (text) summarization.

Most automatic summarization systems involve the generic sequence of steps depicted in Figure 2.2 to achieve the required result. Note, however, that the neatness of the figure is deceptive. Each of the steps is potentially highly complex; their substance tends to vary substantially depending on the specific task to be accomplished. Pre-processing, for instance, might involve nothing more than the tokenization of the documents and sentences within them, or require such multi-faceted processes as deep syntactic analysis and the conversion of the documents to some sort of graph-analytical representation. Likewise, information selection and sentence ordering can be as straightforward as establishing word frequencies, or involve the deduction and subsequent processing of dense syntactic and semantic information. In order to illuminate some of the innumerable subtleties to be considered when composing a summary, automatic or otherwise, of a single *or* multiple sources, the next section attempts to provide a classification of a number of summarization tasks and the approaches to

achieving the required results. For transparency, pre-processing is combined with information selection, as the complexity of these tasks is correlated; higher complexity in the information selection step typically requires more intricate pre-processing.

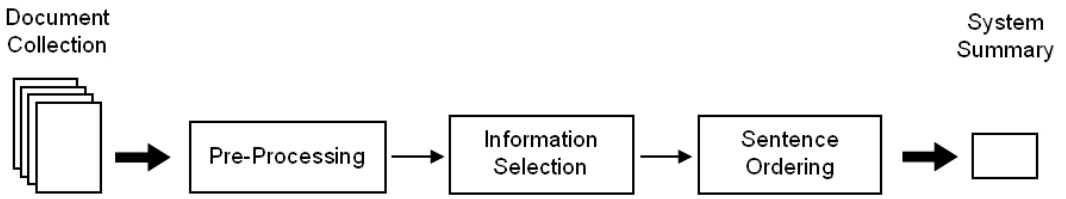

Figure 2.2: The generic process underlying most automatic summarization systems.

2.2 Categorization of Summarization Systems

Beyond ascertaining the type of information to be summarized – in the work to be presented, text from multiple document sources – careful thought must, amongst other things, be given to the required content and depth of processing. For, in its essence, automatic summarization is a complex sequence of tasks drawing from a variety of natural language processing (NLP) tools. As such, to arrive at the optimal series of steps to generate an appropriate summary, it is important to have a clear grasp and understanding of a wide range of aspects relating to the required attributes of the summary to be composed. For instance, highlighting a number of possibilities relating to the summary's depth of understanding, Sparck Jones (1998) characterizes summaries as indicative, informative, critical, or aggregative. The latter two of these classes are typically considered to be well out of the reach of current summarization approaches, as they require more than a content-based understanding of the source text in order to produce a suitable summary. The remaining distinction draws attention to the divergence in detail depending on the original query. While indicative summaries note that a source is about some topic without giving detail, informative ones convey what the text actually states about the topic.

Besides this break-down relating to their depth of understanding, summarization systems can be classified along a multitude of dimensions. The most prominent classifications are briefly described and, where useful, contrasted in the following list, moving from summarization tasks to summarization techniques.

Summarization Tasks

- **Generic *vs.* User-Focused *vs.* Topic-Focused Summarization.**
  This classification emphasizes the various possible informational requirements of the user. While generic summaries contain the salient information of a text, topic-focused summaries extract the most important information given the user's query. In user-focused summarization, in turn, the most important information relevant with respect to some model of the user's needs and interests is summarized.

- **Abstractive *vs.* Extractive Summarization (or Fact *vs.* Text Extraction).**
  Tasks within these categories encompass both the need of the user and the (re)-generation effort involved in creating the summary text. For, in contrast to abstracts, extracts contain sentences, clauses, and/or words from the original text(s) in unmodified form. In other words, the production of extracts does not involve the generation of new text.

- **Single-Document *vs.* Multi-Document Summarization.**
  In principle, this distinction could fall under task *or* approach. It draws attention to the additional informational requirements (and correspondingly processing steps) necessary when summarizing multiple as opposed to a single text source. The reason is that besides content issues, multi-document summarization also needs to take account of and deal with such aspects as "conflicts and contradictions, redundancy, collation, [and] sentence ordering" (Newman et al., 2004). In consequence, single-document summarizers cannot straightforwardly be applied to multi-document summarization tasks.[1]

Summarization Approaches

- **Limited Domain *vs.* Open Domain Summarization.**
  This distinction underlines the specificity of some summarization approaches

[1] A recent investigation by Nenkova and Vanderwende (2005) suggests that one approach to dealing with the problems raised by multi-document summarization, a separate redundancy component, is not in fact necessary, but that direct modeling of multiple occurrences of words in the summarization component is highly effective. In particular, they use a statistical approach on a word level: the probability of a sentence to be included in the summary is taken to be the average probability of the words in the sentence. If words already occurred in previously selected sentences, the probability is adjusted to reflect the probability that the word occurs twice in the summary, i.e., the new probability of the word is the square of the old probability.

to particular domains such as newswire or scientific text. Limited-domain approaches such as argumentative zoning (Teufel, 1999) are tailored (and trained) to the specific needs of a particular subset of texts from a certain domain. Conversely, generic (or open-domain) summarization systems are intended to work irrespective of the domain, field, or topic of the source documents.

- **Shallow *vs.* Deep-Processing Summarization.**
  The degree of linguistic processing employed in the process of generating summaries varies substantially among different summarization systems and tends to depend critically on the task at hand. While some primarily employ very shallow features such as counting word frequencies, tf.idf scores (i.e., term frequency multiplied by inverse document frequency),[2] and sentence position, others employ more linguistic processes such as (partial) parsing, chunking, or semantic relationships. A classic example of shallow processing can be found in Edmundson (1969), while Barzilay (2003)'s information fusion illustrates rather deep linguistic processing based on predicate-argument structures.

- **Knowledge-Based *vs.* Machine-Learning-Based Summarization.**
  The crux of this distinction is that knowledge-based techniques in the sense employed here make no use of machine learning, utilizing a developer's understanding of the processes at hand directly. Examples of knowledge-based approaches include LexRank (Erkan and Radev, 2004), lexical chains (Barzilay and Elhadad, 1997), basic-elements-based summarization (Hovy et al., 2005), frequency-based summarization (Nenkova and Vanderwende, 2005), and feature-weighted sentence extraction (Edmundson, 1969). The latter also forms the basis for many machine-learning approaches, even though the original version uses manually optimized weights. Paralleling research in many other areas of computer science, machine-learning approaches are becoming ever more prominent in automatic summarization. Bayesian approaches (Daumé III and Marcu, 2005a; Kupiec et al., 1995), hidden Markov models (HMMs; Conroy and O'Leary, 2001), support vector machines (SVM; Hirao et al., 2002), and latent semantic analysis (LSA; Gong and Liu, 2001) are examples of just some of the uses of machine-learning techniques as applied to automatic summarization tasks.

Albeit edited to comprise only the most important classifications, what should be

[2]The intuition of this measure is that words that occur often in a document but seldom in the document collection are, in fact, important (Jones, 1972).

clear from this list is that even seemingly simple summaries are fundamentally intricate. They call for a multitude of preparatory decisions to be able to develop a successful methodology, not to mention the determination of the appropriate depth of processing and the like. In the case of automated systems, grasp of these distinctions carries even greater weight, as the summarization system per se is completely naïve.

2.3 Related Systems

The discussion thus far was intentionally kept quite non-technical to facilitate a clear overview of the basic subject and its subtleties. As such, it only referenced some of the basic characteristics of research in automatic summarization. From this section onwards, however, a number of state-of-the-art summarization systems and techniques as they relate to the work carried out for this thesis will be described, assessed, and contrasted.

Using the terminology of the foregoing classification, the approach proposed in chapter 5 of this thesis is a multi-document, informative, generic, deep, open-domain, knowledge-based, extractive summarization system. To underline its novelty and usefulness, the remainder of this chapter considers a fairly diverse set of related (and, at times, tangential) existing approaches, ranging from single- to multi-document summarization. Starting, in this section, with related automatic summarization systems, the survey reviews recent work on determining sub-sentential units such as logical forms, syntactic triples, and discourse units, as well as work on the incorporation of factors beyond informational content in the content selection process. Given their central role in the work at hand, the issues of information selection and sentence ordering will subsequently be explored in separate discussions.

Among the earliest contributions to attempt the integration of informational content with one or more other factors is CLASP (Tucker, 1999; Tucker and Sparck-Jones, 2005), a single-document, generic, open-domain, knowledge-based, informative, extractive summarization system. As a central part of their approach, the authors construct a graph capturing the total of the logical forms in the document. They then use a greedy algorithm to establish the optimal subset of the logical forms to represent the summary. The optimal solution is determined by three separate factors: importance, representativeness, and cohesiveness.[3] The main achievements of their methodology are the inclusion of cohesiveness as a factor in the determination of a summary and the

[3]Text coherence measure are discussed in Section 2.5.

extraction of sub-sentential units. Before CLASP, cohesiveness was only considered as a post-processing step to order the sentences selected in a previous step as opposed to being integrated in the content selection process.

Two approaches building on this graph-theoretic methodology and rooted in a similar outlook (to one another) on the summarization task are Vanderwende et al. (2004), who pursue a multi-document summarization system, and Leskovec et al. (2004), who focus on single document sources. An example of an "input" graph generated by the approach of Leskovec et al. (2004) is shown in Figure 2.3; a comparable illustration based on the approach of Vanderwende et al. (2004) is given in Figure 2.4.

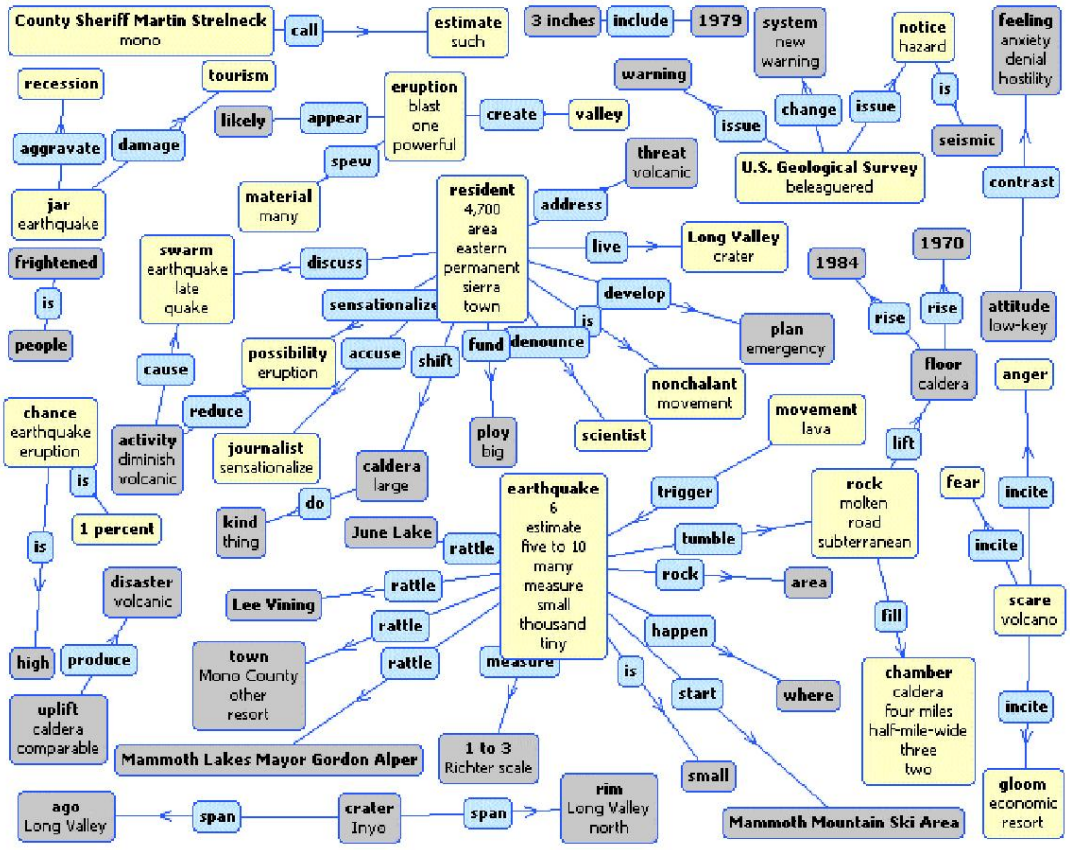

Figure 2.3: Full semantic graph of the document "Long Valley volcano activities." Subject/object nodes indicated by the light color (yellow) nodes in the graph indicate summary nodes. Gray nodes indicate non-summary nodes. (Reproduced from Leskovec et al. (2004), Figure 5.)

The "input" graphs differ insofar as Vanderwende et al. (2004) create a graph in which edges carry the syntactic relationship between nodes, while Leskovec et al.

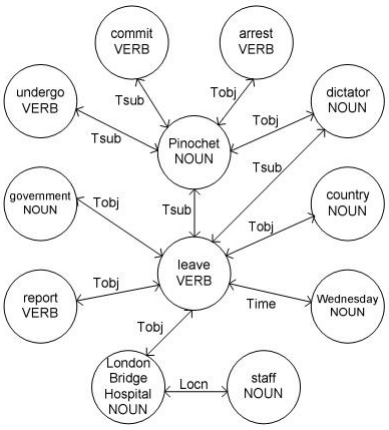

Figure 2.4: Fragment of a document graph for the document representation of Vanderwende et al. (2004). (Reproduced from Vanderwende et al. (2004), Figure 2.)

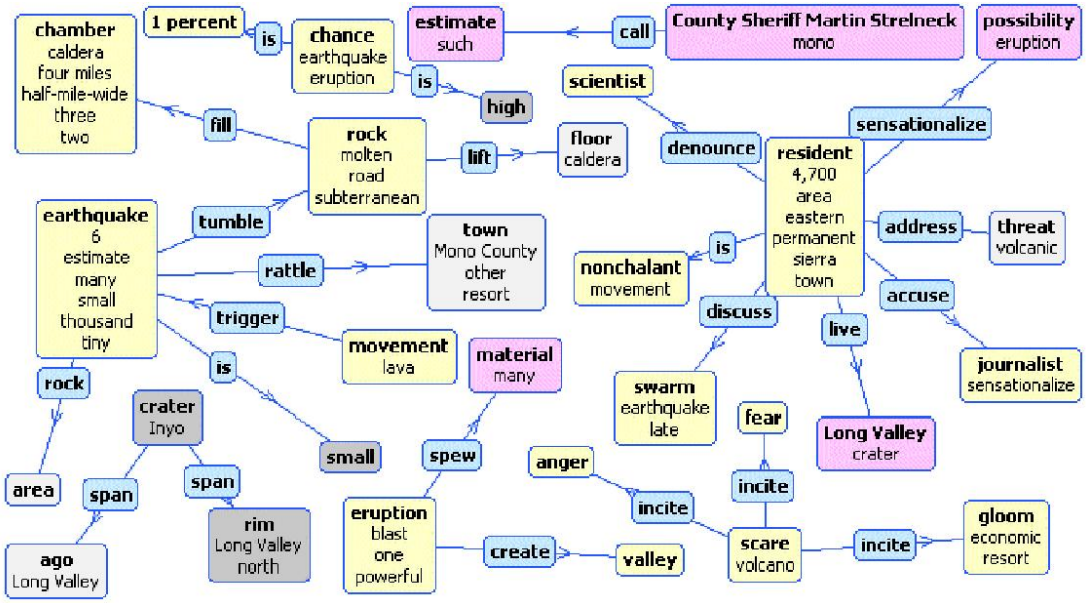

Figure 2.5: Automatically generated summary (semantic graph) from the document "Long Valley volcano activities." (Reproduced from Leskovec et al. (2004), Figure 7.)

(2004) do not attach informational value to edges, calling their graph a semantic graph.[4] Similarly, in the subsequent processing stages, Vanderwende et al. (2004) only use PageRank (Page et al., 1998) to assess importance, while Leskovec et al. (2004) employ a variety of different approaches including PageRank, Hubs and Authorities (Kleinberg, 1999), as well as the size of weakly and strongly connected components. To be more precise, Vanderwende et al. (2004) use PageRank scores directly to identify the most important triples in the document cluster graph, where the triple is the most important node in conjunction with the most important of its neighbors. They extract text fragments from each sentence based on these important triples, where each fragment can either represent an event or an entity. The event fragments are then clustered together based on the event they refer to, whereupon the most informative fragment is selected to represent the cluster. In a final step, the summary is generated by selecting the most important clusters, represented by their fragment, until the byte-length of the summary is reached.

Leskovec et al. (2004), on the other hand, use SVM[5] machine learning on document/ document-summary pairs in order to obtain an optimal classification based on a total of 118 distinct linguistic attributes for each individual node, 14 graph properties from the constructed graph, and further attributes describing approximate discourse structure. Sentences are extracted using a simple decision rule stating that a sentence is included in the summary if at least one of the triples included in the summary triples is present in the sentence. Figure 2.5 depicts the semantic graph *after* the application of their summarization process, i.e., the graph that triggers sentence selection for the summary.

There are two aspects of broad interest to graph-based summarization methods: the creation of the graph and the scoring of the nodes of the graph. In the context of the methodology employed in this thesis, however, it should be noted that these graph-based methods do not use larger syntactic or semantic relations in order to identify similarity of the different nodes in the graph. For example, in Figure 2.3, the sensationalizing of journalists does not explicitly link to the possibility of eruption (lower left of the *resident* node). The main approach of graph-based methods is in the use of

[4]While Leskovec et al. (2004) call their graph a semantic graph, Vanderwende et al. (2004) call their graph a syntactic graph. However, both approaches capture the predicate-argument structure or grammatical relations of the underlying document. As such, both representations represent shallow semantic information.

[5]Support vector machines (SVMs) are a specific set of supervised machine-learning algorithms mainly used for classification and regression analysis. They were originally introduced by Boser et al. (1992).

graph-walking techniques in order to determine the nodes that are most central to the overall graph. In contrast, my approach concentrates on small syntactic and semantic pieces of text that are similar in multiple documents, but does not cover a graph-based centrality measure. From this perspective, graph-based methods represent approaches that use similar information (i.e., syntactic and semantic information) in a significantly different manner to the approach proposed in this thesis.

Another informative area of research is the use of the discourse structure of documents to be summarized in order to determine the most important discourse units. Promising approaches that make use of this notion are Marcu (1998), who uses the depth in the discourse tree and the nucleus-satellite in the structure to obtain the discourse units that are most important, Miike et al. (1994), who use decision trees on the discourse representation, and Thione et al. (2004), whose PALSUMM system combines discourse information with statistical information obtained using MEAD, a public domain multi-document summarization system (Radev et al., 2004). The latter contrasts, for example, with Marcu (1998) who only uses discourse information.

Figure 2.6 provides an illustration of the summarization process using discourse elements. It exemplifies the methodology by Marcu (1998). The small boxes containing numbers only represent text, i.e., a particular discourse element. The boxes of the tree represent particular discourse relations – children in dotted lines are satellite nodes, while the other boxes represent nuclei. The text to be contained in the summary is derived via the propagation of the textual representation of the nucleus child. The partial ordering of the discourse units via this method is $2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6$. Given this ordering, the summary text contains the most important units subject to the length requirement being satisfied.

In conjunction with Leskovec et al. (2004)'s transference of semantic graph scores into triplets and their importance, discourse-based measures illustrate the fact that the units of extraction do not necessarily have to be sentence-sized units. Yet, even though sentence-sized and discourse-unit-sized units are in principle variable in size, they are (still) defined by their characteristics. Leskovec et al. (2004)'s fixed-sized triplets are clearly smaller than most discourse units. To combine the ideas of these approaches, in this thesis, small fixed-sized units are combined to create variable-sized units that are not defined by their syntactic characteristics, but by the semantic characteristics of the underlying documents. As such, the systems presented in this section also illustrate the various decisions to be made with regard to the size of the units extracted from the source documents.

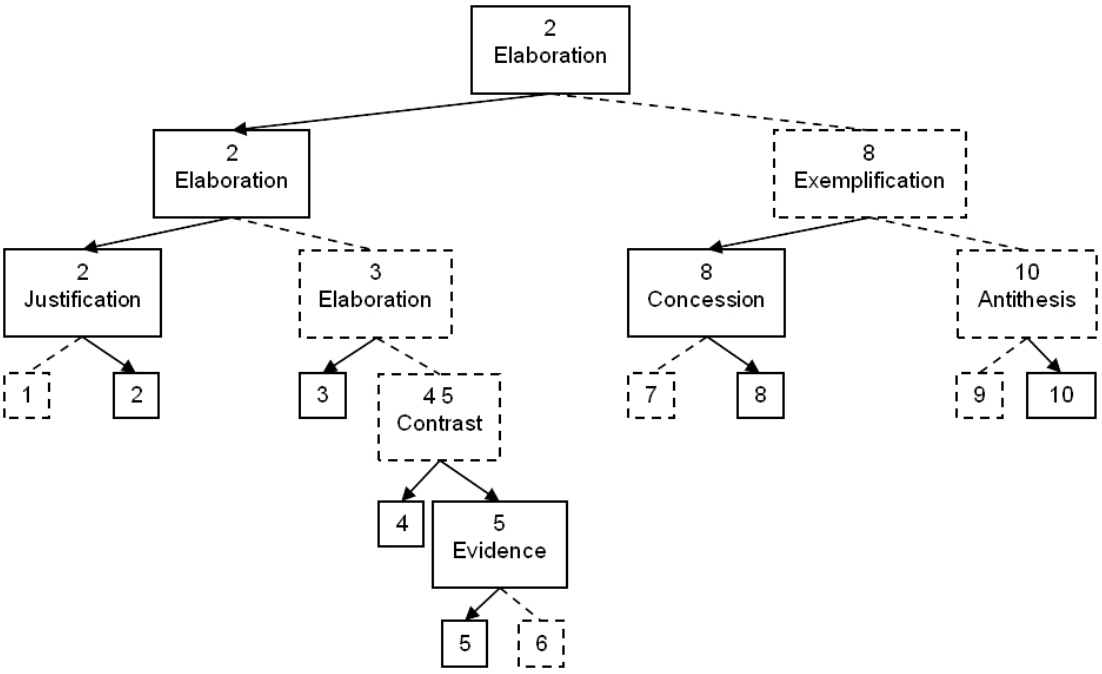

Figure 2.6: An example of rhetorical structure theory and its application in single-document summarization. The figure shows the document representation of a text in discourse units and their interaction based on rhetorical structure theory. Numbers in boxes represent the most important discourse units in the subsumed text. The partial ordering provided by this method is given in the text. (Reproduced from Marcu (1998), Figure 1.)

2.4 Information Selection

In view of the central role of information selection algorithms in the ultimate makeup of a summary, this section delves more deeply into the workings of this component of every summarization system. There are numerous different ways of approaching this problem. The main distinction to be made is the size and composition of the informational units to be selected. In particular, one can (1) directly extract full sentences; (2) extract full sentences on the basis of selected sub-sentence units; and (3) use sub-sentence units to create new sentences. Note that in each approach the size of the informational unit is different. While approaches (1) and (2) have the distinct advantage of syntactic correctness, the merit of approach (3) is that it selects information more accurately, discounting extraneous pieces of information.

A concise way to portray the sentence selection problem (i.e., the information selection process with size of the informational unit being a sentence) is to view it as an algorithmic process by which each sentence in a single document or a collection of documents to be summarized is assigned a score that reflects its importance – in terms of its value for the accurate conveyance of the document's or documents' informational content – within the particular document or collection of documents (as a whole). This perspective is valid for all extractive summarization systems; and if sentence selection was replaced by information selection, it also applies to abstractive summarization. The basis for computing the importance score can be varied based on the particular requirements of a given summarization task, ranging from word-frequency information to discourse information, syntactic, or semantic information. For purposes of comprehensiveness, this section surveys high-level approaches within each of these categories. A detailed exploration of the specific approaches informing the summarization system proposed in this thesis is postponed to Chapter 5.

Approaches based on word frequency comprise some of the earliest attempts to get a handle on sentence selection (Luhn, 1958). However, even quite recent work is based solely on word frequency information. One such contribution is SumBasic (Nenkova and Vanderwende, 2005), which uses the estimated unigram probability of a word occurring in a particular document (given by the number of times the word occurs in the document or document collection divided by the total number of words in the document or document collection) as the basis for generating a summary. For reasons of computational complexity, the authors do not consider all possible combinations, but instead assign each sentence a score given by the product of the probabilities of the

words in the sentence normalized by the number of words in the sentence. They then select the highest scoring sentence to be included in the summary, before updating the probabilities of the words that occurred in the selected sentence by taking into account multiple occurrences in the summary before continuing with the sentence selection process.

The major advantage of word-frequency-based approaches is that they require relatively minor pre-processing, i.e., they typically do not involve syntactic and/or semantic analysis (apart from WordNet synonym-set usage). These approaches therefore tend to be straight-forward and quick. Their main downside is that they are based on an aspect that is only of secondary importance in human summarization: while humans summarize the most important *information*, word-frequency-based approaches summarize the most important *words*. In other words, the notion underlying word-frequency-based approaches is the (weak) inference that if important information is expressed, and is expressed somewhat similarly, words used more frequently should incidentally express more important information. To implement this reasoning, many approaches use stop-word lists to remove frequently used, but usually quite meaningless words (such as "the," "for," or "be"), from the source documents (*cf.* tf.idf and related approaches).

Rather than exploiting word frequencies or the syntactic information of the source documents, (pure) discourse-related approaches are based (exclusively) on the structural composition of the documents to be summarized. Marcu (1998), for example, based his methodology on the assumption that any coherent text consists of a finite number of unique structural building blocks, the composition of which can be used to draw conclusions about their relative (individual) importance. The resulting system correspondingly selects as most salient the sentences closest to the root of the discourse analysis (provided by Rhetorical Structure Theory (RST), a particular flavor of discourse analysis). It should, in this context, be noted that with increasing summary length, the selected sentences are further away from the discourse root. An example of an RST analysis along with the relative importance of the discourse units is provided in Figure 2.6 (see above).

Barzilay and Elhadad (1997) use the number of related concepts (based on WordNet relations) in close proximity to each other as a measure of relative importance for the different concepts. They use lexical chains – a concept describing the distribution of related words in a document as well as their distribution among the sentences in order to determine its main ideas – to find sentences with many high-scoring chains.

The process can be represented in the manner illustrated in Figure 2.7, which is based on the following sample text:

> *Mr. Kenny is the person that invented an anaesthetic machine which uses micro-computers to control the rate at which an anaesthetic is pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anaesthetic into the patient.*

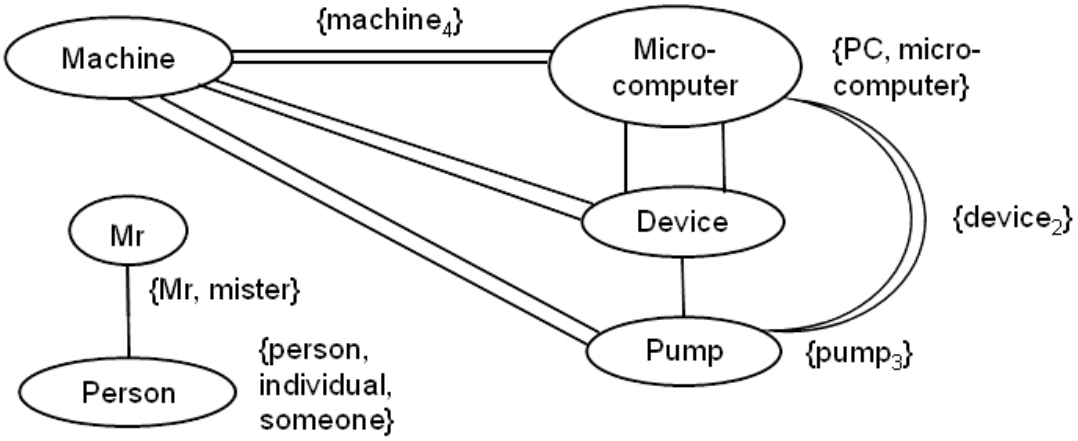

Figure 2.7: An example of the application of lexical chains on the sample passage given in the text. (Reproduced from Barzilay and Elhadad (1997), Figure 4.)

Their approach uses WordNet in order to determine related words within a certain window of sentences, along with dynamic programming, to figure out the optimal composition of the lexical chains. In the figure, which presents the resulting chains, the words in braces denote occurrences of synonyms, while the lines connecting the nodes signify identified connections between the concepts (i.e., words in ovals); multiple lines denote multiple connections between the concepts. The more concepts belong to a chain, the stronger the chain, and the stronger the chain, the more important it is for the summary.

Hence, compared to word-frequency and graph-based techniques, discourse-based approaches utilize different aspects of a document. To be precise, they utilize the grouping of information as opposed to frequency or syntax. The advantage of this variation is that they capture the important information in each section of the text individually. On the downside, however, discourse-based approaches ignore the additional information that can be gained when using the other approaches (such as the frequencies of the words in the discourse units). In addition, authors vary significantly in the

kinds of discourse structure they use, while word-frequency and graph-based information remains relatively constant. Correspondingly, discourse-based approaches are very good if the author(s) structured the text well.

Moving on to syntactically-motivated sentence extraction (which typically also accounts for word frequencies), the work by (Tucker, 1999; Tucker and Sparck-Jones, 2005) is a relevant, recent example of the type of processing common in this category. As indicated above, they use syntactic analyses in order to construct a graphical representation of the "input" document. The most interesting aspect of their work, however, is that they model their information-selection process not only based on unit importance (in the abstract notion that some unit is globally important irrespective of which other units are selected), but also consider representativeness and cohesiveness to be important factors in summary creation. This aspect is different to most other graph-based methods, including Vanderwende et al. (2004), who do not include other factors. The main problem with Tucker and Sparck-Jones (2005) is the use of a fixed manual weighting between the three attributes (importance, representativeness, and cohesiveness). The difficulty with this design choice is the assumption that all information is selected in the same manner. I would, however, argue that few humans create summaries in this manner. Instead, they first select the most important information and *then* select information that connects the most important information selected previously; they subsequently revise the information selected in order to fit external constraints such as length.

An approach in a similar spirit, but using semantic information to assign importance scores is that by Daumé III et al. (2009). It is based on a vine-growth model, which picks a sentence to be summarized and then, in a step-wise procedure, decides whether to select components within the sentence to be included in the summary, or whether to start a new sentence. More broadly, the system uses a machine-learning algorithm that incorporates its own incorrect predictions into the prediction process (a meta-algorithm called SEARN), thus considering uncertainty at learning time as opposed to search time. The decision criterion to be optimized by the algorithm is the ROUGE score. The combination of these attributes in their search-algorithm results in a performance that nearly equals an oracle score for 100-word summaries, but does significantly worse for 250-word summaries. The algorithm is exemplified in Figure 2.8. Assume that the algorithm decided that a new sentence should be selected and that the sentence in the figure is that sentence, i.e., "The man ate a big sandwich with pickles." The algorithm starts out with an empty sentence and decides whether to start a new

sentence or select one or more words from the selected sentence. In the present case, it decides to select "ate." In the next stage, the algorithm once again needs to decide whether a new sentence should be started or additional words should be selected. This process continues until the algorithm has selected "The man ate a sandwich." At this point the algorithm determines that a new sentence should be selected, and the process starts again.

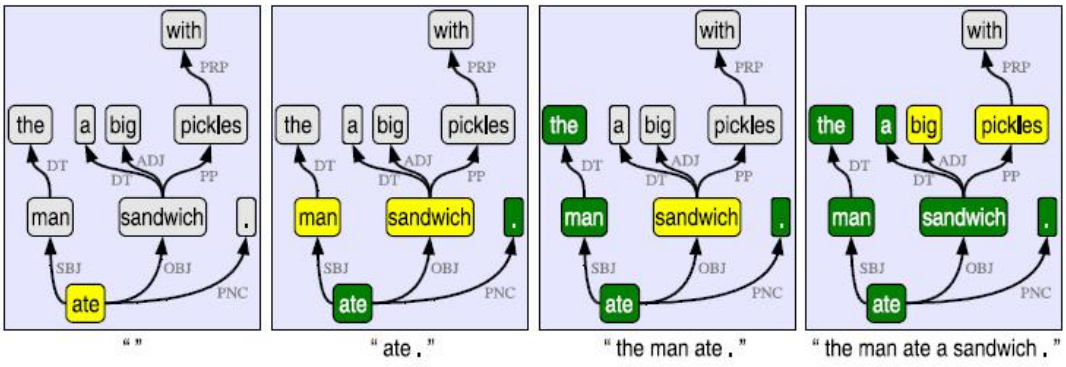

Figure 2.8: An example of using SEARN for the summarization process. The sentence for which the dependency (syntactic) parse is shown is: "The man ate a big sandwich with pickles." (Reproduced from Daumé III et al. (2009), Figure 6.)

The interesting aspect about this system is the selection of variable amounts of information from each sentence selected for the summary. In particular, the system shows that iteratively selecting words to be included in the summary works quite well. Note, however, that much of the complexity of the approach is hidden in the selection of the parameters for their machine-learning algorithm, SEARN. For the work presented in this thesis, the central concept is the step-wise selection of information, though the idea is applied in a somewhat different manner to the end of constructing a (fully) automatic evaluation scheme.

2.5 Sentence Ordering

While typically not viewed as the most important aspect of summarization, sentence ordering is a necessary step to create an easily understandable summary (text). If the information in the summary is presented in a confusing or misleading manner, then irrespective of the quality of the selected information, the resulting summary is not readily helpful to a human user. In more cases than not, the ordering methodology

is specific to each particular system. Accordingly, this section restricts itself to delineating the basic task and highlighting the conceptual notions and features common across the various schemes. A detailed discussion of the specific attributes utilized in the proposed system is presented in Chapter 6.

Paralleling the discussion on information selection, sentence ordering can be regarded as a subfield of information ordering, the ordering of any nuggets of information, not only full sentences. While this task is relevant to summarization, it is not the only field in which information ordering is important. Amongst other things, natural language generation systems also need to consider how information is presented.

In general, sentence ordering is a difficult task because of the nature of linguistic coherence. Coherence (in linguistics) is what makes a text semantically meaningful (Lalitha Devi et al., 2009), i.e., a text rather than a collection of unrelated (seemingly random) statements. There are two main aspects to coherence: (1) the purely linguistic elements subsumed under the notion of cohesion (the grammatical and lexical relationships within a text); and (2) the presuppositions and implications associated with general world knowledge.

From this theoretical perspective, sentence ordering would need general world knowledge to achieve truly coherent text. However, one of the standard assumptions in sentence ordering research is that there is a single correct coherent ordering for sentences (selected given a specific task). Such an assumption is correct in tasks where the original text is a coherent text (e.g., the identification of the original ordering in documents containing re-orderings of the original (Barzilay and Lapata, 2008)). However, for the purposes of the sentence ordering task in summarization the assumption does not generally hold true because there usually does not exist an original ordering for the selected sentences. As such, in limited-domain summarization systems, template-based approaches (Radev and McKeown, 1998) use domain-specific knowledge regarding the information and ordering to occur in the summary (e.g., timeline, event, and the like). Such approaches explicitly encode general world knowledge for the particular domain in order to generate meaningful summaries. In the context of open-domain sentence ordering systems, however, this tactic is impractical as it would entail encoding this sort of general world knowledge for each and every domain. Therefore open-domain summarization systems need to bypass the general world knowledge aspect by way of a variety of approaches.

The basic idea driving work in this area is the notion that similar (and/or related) information should be grouped together. The standard solution to this problem tends

to be a simple grouping rule based on the sentences' cosine similarity (Bollegala et al., 2005). Alternative approaches involve temporal ordering (Bollegala et al., 2006), ordering based on the topic of the sentences (as determined in numerous ways) (Barzilay et al., 2002), ordering of the sentences in the same way as in the underlying document or document collection (Bollegala et al., 2006; Barzilay et al., 2002), and ordering based on the publication date of the document from which the sentence(s) were extracted (Barzilay et al., 2002). In the case of single-document summarization the ordering based on the ordering in the original source document tends to be the most appropriate ordering (Barzilay et al., 2002). In multi-document summarization, however, this approach is not feasible due to conflicts in the ordering between the various source documents. Other approaches tend to be better suited for particular domains of summarization, i.e., the underlying source documents. For example, ordering based on the publication date of the source document works well for a document collection presenting a timeline of an event (say, a war and the different events leading up to its resolution), though it would not be optimal in the case of some philosophical or emotional discussion. In the same vein, it should be clear that ordering by topic (only) would not necessarily be the best option for documents presenting a timeline, again, say, a war and its resolution, because such summarization would group political/diplomatic efforts, while military efforts would present a different section of the summary document. The result would be a loss of the timeline associated with the described events and the intrinsic connection between the conflict and the events leading up to its resolution described in the sources. In short, the motivation, benefit(s), and shortcoming(s) of the various approaches to sentence ordering are closely related to the summarization task. The general overview of the problem provides the motivation for the research in Chapter 6.

2.6 Evaluation of Automatic Summarization Systems

Having constructed an automatic summarization system, a natural final step is the assessment (or evaluation) of its performance relative to existing schemes and techniques to gauge its merits and weaknesses. This procedure is essential for the progress of research in NLP. Indeed, in most areas within NLP (such as part-of-speech tagging, parsing, and the like), accepted evaluation methods are readily available, for instance, using gold-standards, which compare system output to the correct output as given by the gold-standard. In summarization research, however, this is not necessarily the case.

The main problem in this context is the fact that there is no such thing as a "correct" summary. Even humans do not tend to agree which information should be contained in a given summary (Teufel and van Halteren, 2003). As a consequence of these difficulties, much effort has been directed towards finding better evaluation methods, in the sense of facilitating better comparability between approaches. This section provides an overview of the various possible approaches to evaluating automatic summarization systems, and describes and contrasts some of the most commonly used evaluation systems in recent years.

2.6.1 Widespread Approaches to Evaluating Automatic Summarization Systems

The evaluation of automatic summarization systems is a difficult problem and, in itself, an active field of research. The major forum for the evaluation of summarization systems are the Document Understanding Conferences (DUCs), incorporated into the Text Analysis Conference (TAC) since 2008. The first DUC took place in 2001. Each year, the conference invites submissions for a different task – the foci in recent years included single-document summarization, generic and user-focused multi-document summarization, as well as the evaluation of summaries.

In general, the task can be approached from two directions: intrinsically or extrinsically (Sparck Jones and Galliers, 1996). Broadly speaking, intrinsic methods evaluate the summary against certain inherent measurable characteristics, while extrinsic methods measure the usefulness of the summary for some task. Note that, depending on the particular evaluation task, both methods can involve manual and automatic techniques.

Intrinsic Evaluation. Evaluation methods in this category (directly) measure the peer summary's performance on specific, pre-defined attributes, which the evaluator considers important constituents of high-quality summaries. To this end, the effects of the two main tasks of summarization, information selection and text production, are typically considered separately. Information selection can be captured by a variety of measures of informational content such as ROUGE (Lin, 2004) or the Pyramid (Nenkova and Passanneau, 2004) evaluation scheme (see below), while text production can, amongst other things, be assessed by measures for readability, grammaticality, and structure, such as the 5-point manual evaluation scales used in the DUCs. Attributes such as readability and grammaticality can usually be assessed rather quickly and consistently. The assessment of informational content, via naïve methods such as the as-

sessment by a human evaluator on a 5-point scale, on the other hand, is time-consuming and unreliable (Nenkova and Passanneau, 2004). Much of the current research therefore concentrates on this aspect of intrinsic evaluation.

Extrinsic Evaluation. Extrinsic methods, also referred to as evaluation "in use," strive to measure the usefulness of the summaries produced by a given system in a complex (e.g., real-world) task. An example of an applicable task is the relevance assessment of documents in the context of information retrieval (Daumé III and Marcu, 2005b). The idea is that, using summarization systems, one obtains decisions of at least the same quality in a smaller amount of time than when using the full documents. In other words, summarization might allow the processing of more material in the same (or a lesser) amount of time. The obvious drawback of this evaluation technique is that the results do not highlight specific strengths and weaknesses of the underlying summarization system, but merely report the (aggregate) performance on the given task, which may not reveal underlying problems.

While the preceding paragraphs delineated the various general aspects to be considered when evaluating summaries, the remainder of this section contrasts a number of (recent) methods for evaluating the informational content of summaries. Note, in particular, the Pyramid evaluation scheme, which constitutes the foundation for much of the work presented in this thesis.

2.6.2 Recent Evaluation Methods to Assess the Informational Content of Automatically Generated Summaries

In the early days, the evaluation of summarization systems was often based on a single human reference sample. Recent findings, however, suggest that human summaries tend to exhibit significant variation (Teufel and van Halteren, 2003). The natural implication is that a single human summary is quite unlikely to be a reliable reference to gauge the true quality of the information contained in an automatically generated summary. In consequence, as illustrated in Figure 2.9, modern approaches – particularly those surveyed in this section – commonly rely on multiple human reference summaries.

As exemplified by many of the contributions to DUC 2001, manual evaluation tends to produce quite variable results when used to assess the informational content of automatically generated summaries, calling into question the results' dependability. This has inspired a lot of research, not only with the objective of obtaining more re-

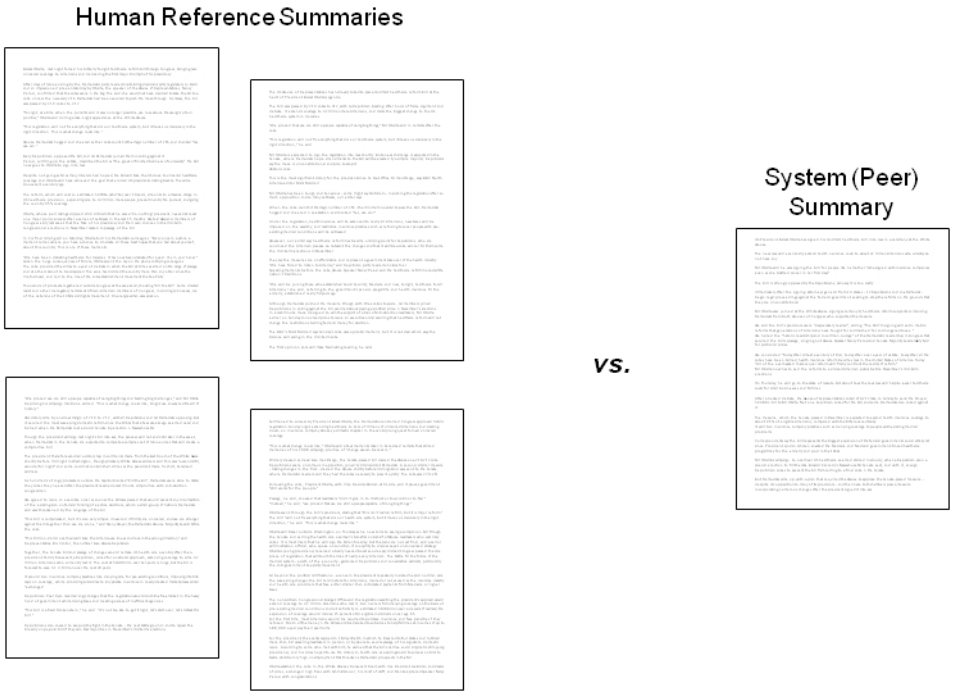

Figure 2.9: Modern Summarization Evaluation. Modern approaches to evaluating the informational content of automatically generated summaries compare a single system (or peer) summary against multiple human reference summaries.

liable human judgments, but quite a bit of effort has been expended into constructing automated evaluation techniques (e.g., Teufel and van Halteren, 2003; van Halteren and Teufel, 2004; Nenkova and Passanneau, 2004; Radev et al., 2000, 2003; Lin and Hovy, 2003; Hovy et al., 2005; Passonneau and Nenkova, 2003). On account of their efficiency, this section reviews some of the state-of-the-art automatic techniques.

2.6.2.1 ROUGE

ROUGE (Lin, 2004) is the current, predominant automatic evaluation method for assessing the performance of automatic summarization systems. It is based on n-gram co-occurrence statistics; the concept is rooted in the idea motivating BLEU (Papineni et al., 2002), a successful evaluation method in the machine translation community. While a few other automatic systems in the same spirit, for instance, cosine similarity and longest common subsequence (Saggion et al., 2002) have been proposed, they have not been correlated with human judgments. The most commonly used versions of ROUGE are ROUGE-N and ROUGE-S.

ROUGE-N refers to a technique involving the n-gram recall between a candidate

summary and a set of reference summaries, or in formal terms:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

where n denotes the length of the n-gram, and $Count_{match}(gram_n)$ signifies the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. The problem with this version of ROUGE is that ROUGE-N is very sensitive to any differences in phrasing across summaries, however small. For example, the replacement of determiners or insertion of adjectives to otherwise plain sentences causes substantial deterioration in the ROUGE-N score for $N > 1$, since ROUGE-1 only captures unigram overlap between the reference and system summaries. Figure 2.10 provides a simple illustration of the problem. If the first sentence is taken to be the reference sentence, the second and third sentence share the same number of bi-grams. It should, however, be unambiguous that the second sentence is more similar to the reference sentence than is the third.

Alice called Bob.
Alice called **lazy** Bob.
Alice called **lazy Fred**.

Figure 2.10: A simple illustration of the deterioration of the performance of ROUGE-N because of the insertion of adjectives.

One of the most common alternative variants of ROUGE, ROUGE-S, is based on the skip bi-gram co-occurrence statistic. This approach allows for gaps of arbitrary length between the first and second word in a bi-gram, though the size of the gap can be adjusted in order to avoid spurious skip bi-grams such as "the the." The advantage of this variant compared to Rouge-N is its ability to better recognize the similarities in slightly divergent surface realizations such as inserted adjectives or even prepositional phrases. Lin (2004) found that, among ROUGE methods, ROUGE-1 (unigram matching), ROUGE-2 (bigram matching), ROUGE-S4 (skip bi-gram matching with up to 4 words between the words in the bi-gram), and ROUGE-S9 (skip bi-gram matching with up to 9 words between the words in the bi-gram) provided the best correlation with human judgments. When applied to the example in Figure 2.10, ROUGE-S1 entails that the second sentence has two skip bi-grams in common with the reference (first) sentence, while the third sentence only shares a single skip bi-gram. Because

of the additional common skip bi-gram, this method now correctly concludes that the second sentence is more similar to the reference sentence than the third sentence.

In general, the advantages of the ROUGE evaluation methods are that they are fully automated, rely on multiple reference summaries, and involve the aggregation of individual comparisons as opposed to a single score such as the cosine similarity between documents. Their main disadvantage is that they largely rely on surface similarities as opposed to syntactic and semantic similarities, which more fully captures the meaning of the text as opposed to the words in the text.

2.6.2.2 Factoids

One of the earliest (manual) evaluation techniques to base its assessment on more than one reference summary is the Factoids approach introduced by van Halteren and Teufel (2003). The characteristic feature of this approach is the emphasis on comparing a system summary's informational content to reference summaries as opposed to assessing string similarity only, as was the case in earlier approaches. According to this approach, human annotators first identify units of information, Factoids, in the reference summaries. They then annotate the peer summary with reference to the factoids from the reference summaries. Thus, the size and composition of the individual units of information within the system summary (the Factoids) is based on the units identified in the reference summaries. As a result, a single Factoid can range from being represented by a single word to comprising an entire sentence. Note that their definition of Factoid is very narrow; the two text fragments "was killed" and "was shot dead" result in three factoids containing the fact that there was an attack, that someone was killed, and that a gun was used in the attack. As detailed below, this classification contrasts with the Pyramid evaluation method, which would tend to group the fragments as having roughly the same meaning.

Given a list of Factoids identified in a peer summary, van Halteren and Teufel (2003) put forward two different methods to obtain a score for its performance:

- Consensus Summary. The score according to this technique is based on the overlap between the Factoids in the system-generated summary and the Factoids in a consensus summary, which is created based on the most frequent Factoids from the reference summaries.

- Frequency-Weighted Factoid Score. In this scenario, the Factoids in the system summary are weighted by the number of times they occur in the reference sum-

maries and the score is the sum of the weighted scores of the Factoids. Note that the use of frequency weights renders this method very similar to the Pyramid evaluation method described in the next section. The main difference is the atomicity of Factoids as compared to Pyramid summary content units.

For clarity, Figure 2.11 presents an example of Factoid annotation. While FA10 and FA40 are present in both sentences (A and B), FA20 is only present in sentence B. If additional information were known (Sentence C), then FP20, FP21, FP24, FP25, and FP26 might represent additional Factoids, depending on the Factoids present in other summaries.

Representative Text Fragments from reference summaries
A: The victim was killed.
B: The victim was shot dead.

Factoids
FA10 There was an attack (in both sentences).
FA40 The victim died (in both sentences).
FA20 A gun was used (in B only).

Additional Information
C: The police have arrested a white Dutchman.

Factoids:
FP20 A suspect was arrested.
FP21 The arrest was carried out by the police.
FP24 The suspect is white.
FP25 The suspect is Dutch.
FP26 The suspect is male.

Figure 2.11: Examples of the factoid annotation scheme. The identifiers at the beginning of the lines identify the respective Factoids. (Reproduced from van Halteren and Teufel (2003) and Teufel and van Halteren (2003).)

Nonetheless, the theoretical analysis by van Halteren and Teufel (2003) shows that even this fairly straightforward approach is not without conceptual problems, as they find that a stable set of units of informational content and their relative importance

requires approximately 40 reference summaries. In this regard, Nenkova and Passanneau (2004)'s Pyramid annotation scheme would seem to be superior as they obtain stable results with as few as 5 reference summaries. Note that, in sharp contrast to ROUGE, this evaluation method is manual, which represents both an advantage and a disadvantage. On the positive side, the manual labor enables the method to consider more fine-grained informational units and facilitates the consideration of semantic relationships, which are not used by ROUGE. On the downside, however, this evaluation method requires significant manual effort in order to obtain scores for the system summaries.

2.6.2.3 Pyramid Evaluation Scheme

The Pyramid evaluation method (Nenkova and Passanneau, 2004; Nenkova et al., 2007), too, is a manual scheme to assess the informational content of peer summaries. It is rooted in four observations about the evaluation of summaries:

- **Human Variation.**
  As different people assign importance to different informational content, a single reference summary does not suffice to determine the quality of system-generated summaries.

- **Analysis Granularity.**
  As the information in sentences can overlap, the most appropriate unit of analysis is not necessarily obvious. Even so, given its recurrent incidence, an evaluation method should account for and cope with this fact.

- **Semantic Equivalence.**
  The same information can be expressed in many different ways, using different wording. In consequence, an evaluation method should be as independent of the actual wording of the information as possible, i.e., it should be based on the semantic content as opposed to the surface realization of the semantic information.

- **Extracts *vs.* Abstracts.**
  With the development of more advanced non-extractive summarizers, it becomes ever more important to incorporate semantic equivalences.

Taking these details into account, the pieces of information that form a unit of (common semantic) content according to this approach are called semantic content

units (SCUs). An SCU is an informational unit the size and content of which is determined via the reference summaries. If all (human) reference summaries agree about the details of some information, then the SCU contains all this information. If, however, one or more summaries contain only a part of this information, then the information is split into two SCUs; one containing the common information between the summaries, and the other the (additional) information contained in only some of the reference summaries. Individual occurrences of an SCU in a summary are called contributors. Note that a contributor can contain multiple parts, as the contained text does not have to be contiguous. The importance (or weight) of a given SCU is determined by the number of contributors it contains. In the above case, this would mean that the SCU containing the common information has a higher weight than the SCU containing the additional information. The name of this evaluation method derives from the fact that the weighted SCUs can, figuratively speaking, be "sorted" into the shape of a pyramid: the SCU occurring most often, thus having been assigned the highest weight, is at the top of the pyramid, while the least frequently occurring SCUs are at the bottom (Figure 2.12).

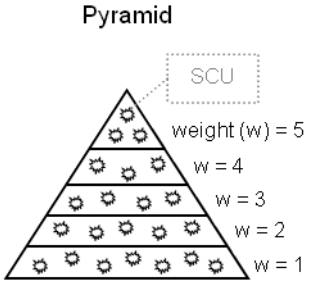

Figure 2.12: The Pyramid. The arrangement of the SCUs (bubbles) into a pyramid according to their relative weights.

The scoring of summaries is achieved by dividing the weighted sum of the SCUs contained in a summary by the maximum weighted sum that can be achieved using the same number of SCUs. In mathematical terms, suppose a pyramid has n tiers, with T_n being the top tier and T_1 being the bottom tier. The weight of the SCUs in tier i is i, because the relevant SCUs occur in i summaries. $|T_i|$ therefore denotes the number of SCUs in tier i. Let D_i be the number of SCUs in the summary that are in tier i. Then the total weight of the summary is $D = \sum_{i=1}^{n} i \cdot D_i$. The optimal score that can be achieved by a summary containing Z SCUs, in turn, is given by

$$D_{Max} = \sum_{i=j+1}^{n} i \cdot |T_i| + j \cdot (Z - \sum_{i=j+1}^{n} |T_i|),$$

where $j = max_i(\sum_{t=i}^{n} |T_t \geq Z)$. The summary score is then given by D/D_{Max}.

Figure 2.13 provides an example of the Pyramid annotation scheme. Each sentence is identified by a letter and number, the letter specifying the summary the sentence came from, and the number specifying the relative position in that summary. The summary content unit SCU1 has a weight of 6, meaning that there are 6 contributors, i.e., 6 instances of the SCU in different reference summaries. The text following the colon is the label of the SCU, a natural language text stating the semantic content of the SCU.

In their final incarnations, the Factoid and Pyramid evaluation schemes are very similar from a procedural point of view. The main difference is their respective take on what constitutes an informational unit. While the Factoid evaluation scheme takes a very narrow view of what represents the same semantic content, the Pyramid evaluation scheme allows for variation in content while still classing it as representing the same semantic content. This difference is illustrated by the distinction between "killed" and "shot dead," which results in three Factoids (*cf.* Figure 2.11), while the Pyramid scheme does not distinguish between "hiring," "recruitment" and "leaving," all of which might have different connotations as regards Lopez's reason(s) for leaving GM and whether or not he accepted the job at VW prior to leaving GM (cf. Figure 2.13).

The advantage of using the Pyramid evaluation scheme as a basis for the automation of the evaluation of informational content (accounting for semantic content as opposed to ROUGE) is the broader scope of the content in the content units. The use of Factoid analysis would require much more accurate semantic analysis, which is problematic because of the fine-grained semantic difference that would need to be captured. What is more, a lot more annotations are available for the Pyramid evaluation scheme because of its use in the official evaluations of the DUCs. Thus, this thesis endeavors to improve the evaluation of summarization systems by automating a manual evaluation metric, thereby providing ease of evaluation along with the accuracy of manual methodologies.

Representative Sample Text

A1. The industrial espionage case involving GM and VW began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.

B3. However, he left GM for VW under circumstances, which along with ensuing events, were described by a German judge as "potentially the biggest-ever case of industrial espionage".

C6. He left GM for VW in March 1993.

D6. The issue stems from the alleged recruitment of GM's eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.

E1.On March 16, 1993, with Japanese car import quotas to Europe expiring in two years, renowned cost-cutter, Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing and Production director.

F3.In March 1993, Lopez and seven other GM executives moved to VW overnight.

SCU1 (weight=6): Lopez left GM for VW

Contributors:

A1. the hiring of Jose Ignacio Lopez, an employee of GM ... by VW

B3. he left GM for VW

C6. He left GM for VW

D6. recruitment of GM's ... Jose Ignacio Lopez

E1. Agnacio Lopez De Arriortua, left his job ... at General Motor's Opel ... to become Volkswagen's ... director

F3. Lopez ... GM ... moved to VW

Figure 2.13: An example of the Pyramid annotation scheme. The letter-number combination at the beginning of the line indicate the summary (letter) and sentence in the summary (number), from which the text in the line is taken. (Reproduced from Nenkova et al. (2007).)

2.7 An Overview of the Thesis

In view of this survey of existing systems and techniques, Figure 2.14 provides an overview of the different elements proposed in the remainder of this thesis and their relation to one another. It illustrates the summarization and evaluation processes from the original source documents, to the generation of a system summary, and all the way to the derivation of a score for the quality of a summary's informational content. In particular, considering first the evaluation process, Chapter 3 details my approach to automating the matching of SCUs into the peer summaries, which constitutes a partial automation of the (original) Pyramid evaluation scheme. Chapter 4 then automates the generation of the SCUs and the associated pyramid. Combined, Chapters 3 and 4 fully automate the Pyramid evaluation scheme. Shifting gears to summary generation, Chapter 5 applies the techniques developed to create a pyramid for the evaluation process to the selection of relevant information to create a summary. Finally, Chapter 6 investigates the problem of ordering the sentences selected by the tools proposed in Chapter 5.

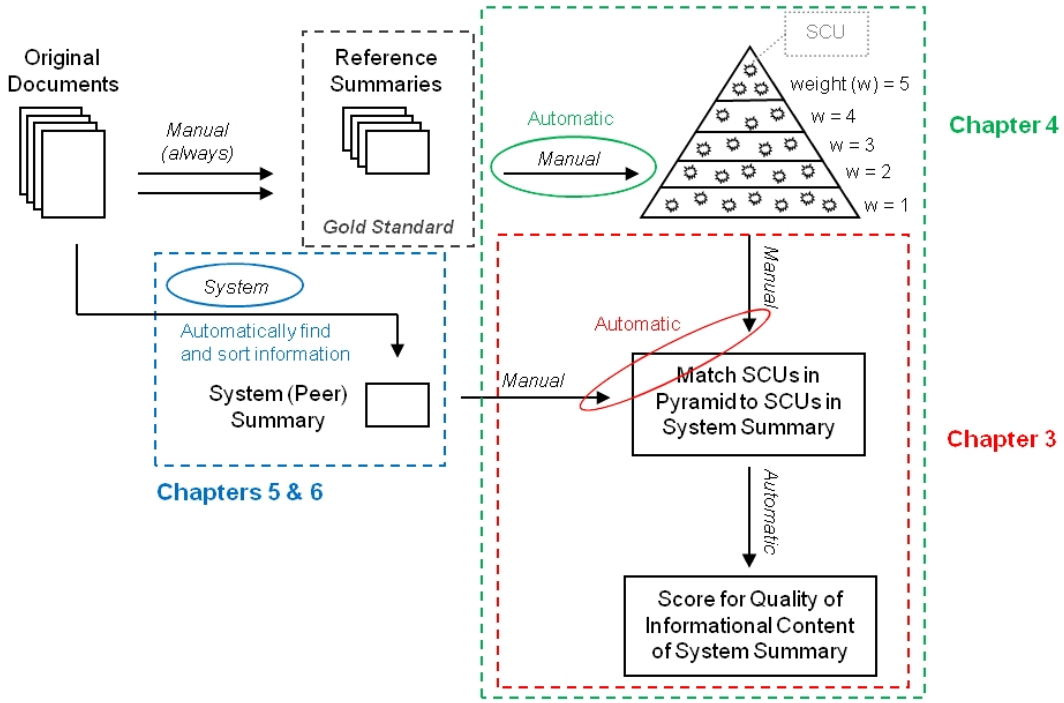

Figure 2.14: A graphical overview of the research presented in this thesis. Note in particular the relationship between the summarization and evaluation steps and the chapters in this thesis.

Chapter 3

Partial Automation of the Pyramid Evaluation Method

3.1 Introduction

The evaluation of automatic summarization systems, a complex but crucial part of their development, is at present predominantly achieved via manual evaluation schemes. One of the most obvious and serious disadvantages of this state of affairs is that the time and human effort required is enormous. A seemingly straightforward remedy to this problem is the construction of *automatic* evaluation procedures. However, since the summarization task allows for a considerable amount of personal and artistic freedom, assessing the quality of a given summary – whether based on a single or a collection of documents – is far from trivial. To name only three complicating factors:

- a given piece of information can be expressed in a multitude of (different but equally expressive) ways;
- the content to be conveyed can be ordered in a considerable number of ways; and
- different authors may attach importance to vastly different content and/or details.

Correspondingly, although quite a bit of research has been devoted to this problem, progress has so far been rather limited. To allow for some variation in summary content, modern evaluation measures are no longer based on a single human reference summary, but involve a collection of summaries produced by different individuals. Even so, the most widely used evaluation scheme to assess the informational content of automatically generated summaries is a fully manual evaluation measure – the two-step Pyramid method (Nenkova and Passanneau, 2004).

As the purpose of this chapter is to present a methodology for the partial automation of this technique, let me briefly reiterate its main attributes (*cf.* Chapter 2). In the first step, like pieces of information are deduced from the reference summaries and assigned a frequency score. This score determines the relative importance of the piece of information. The ensuing ranking can be visualized as a pyramid. In the second step, the evaluator determines which of the various pieces of information can be identified in the automatically generated summary. The final score for a summary derives from the relative frequencies of the units of information in it relative to the maximum possible score for the number of summary content units (SCUs) it contains. Figure 3.1 illustrates the scheme using an example. The sample summary contains three SCUs. Looking them up in the associated pyramid, SCU1 has a weight of 5, SCU2 one of 3,

and SCU3 a weight of 1. Therefore, the sum of the SCUs' weights is 9. The maximum weight possible with 3 SCUs is 15, since 3 of the SCUs in the pyramid have a weight of 5. Thus, the score of the sample peer summary is 9/15, or 0.6.

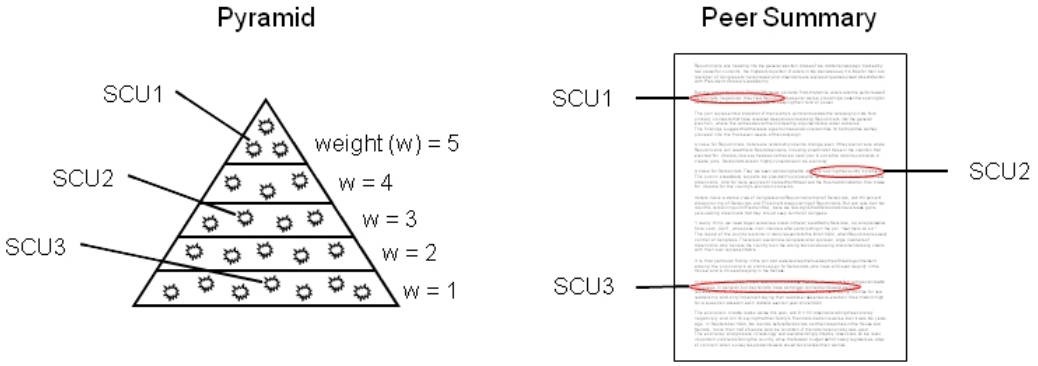

Figure 3.1: An example of the application of the Pyramid evaluation scheme. The sum of the SCUs' weights divided by the maximum weight possible results in a score of 0.6.

With this background, the methodology proposed in this chapter partially automates the Pyramid scheme via an algorithm that automatically matches the pieces of information in a given (manually constructed) pyramid into the associated peer summaries (i.e., Step 2). While the ultimate objective is the full automation of the Pyramid scheme – presented in Chapter 4 – the automation of the second step in and of itself yields a number of benefits. Amongst other things, the proposed procedure:

- facilitates the identification of a number of linguistic, syntactic, and semantic resources more widely useful for determining like information;
- enables the utilization of evaluation material outside the main evaluation event. Thus far, the comparison of results at a later date to the results in the main evaluation event was problematic because of the human effort expended in the evaluation; there may arise a human bias if different human annotators are used across events; and
- can constitute a baseline for human evaluators, who can correct the automatic output to obtain human levels of correctness in a smaller amount of time than would be required for a fully manual evaluation.

To this end, the remainder of this chapter is organized as follows. Section 3.2 describes previous approaches to automating the evaluation of system generated summaries, which explicitly inform the proposed methodology. Section 3.3 subsequently

introduces the architecture and the conceptualization underlying its implementation. Rooted in this more general overview, Section 3.4 outlines the relevant pre-processing steps, followed by a comprehensive overview of the main processing steps in Section 3.5. To gauge the quality of the procedure, Section 3.6 presents a set of experiments to compare the automated approach to the original Pyramid scheme. Section 3.7 highlights the strengths and weaknesses of the proposed system by way of several examples. Section 3.8 concludes the chapter with a brief discussion of the main results.

3.2 Related Work

As detailed in Chapter 2, the most prominent approaches to evaluating a system summary's informational content are the Pyramid evaluation method (Nenkova and Passanneau, 2004), Factoid analysis (van Halteren and Teufel, 2003), and ROUGE (Lin and Hovy, 2003). In their essence, each of these methods is an intrinsic technique focusing on assessing a summary's informational content irrespective of other properties such as grammaticality or structure. As discussed, the Pyramid scheme and Factoid analysis are very similar in their underlying approach. Both are fully manual, rely on multiple reference summaries, and capture semantic content as opposed to surface similarities. In contrast, ROUGE – the only fully automatic approach among the given set – relies on surface similarities between system and reference summaries in the form of n-gram overlap, though it is also based on multiple reference summaries.

An alternative manual approach, utility-based evaluation (Radev et al., 2000), follows a very different approach. The method is designed for the evaluation of extractive summarization systems only. It does *not* rely on reference summaries. Instead, it uses human annotation of importance of the sentences in the *original* documents. Each sentence in the original document is assigned an importance score along with subsumption relations between sentences, i.e., whether a sentence contains (all) relevant information also given in another sentence. While this approach allows for relatively straightforward evaluation of a new extractive summary, the human effort in annotating all sentences in the original document collection is enormous.

A key advantage of utility-based evaluation compared to other manual methods is that one can easily compare the performance of new summarization systems to old systems using the same data. Its major drawback is that it assumes that the summary being evaluated is extractive. Many recent summarization systems, however, do not conform to this assumption, as they modify the sentences they extract. Information

Fusion (Barzilay, 2003), for example, substitutes noun phrases based on a number of criteria. Other systems employ pre- or post-processing steps to simplify selected sentences (e.g., Vanderwende et al., 2007; Jing, 2000; Knight and Marcu, 2002). As soon as the sentences in the summary are no longer replicated directly from the source document(s), the basis of the utility-based evaluation approach ceases to apply. In sum, although utility-based evaluation provides a very different approach to evaluating the informational content of peer summaries, its focus on extractive summaries and manual annotation severely limit its usefulness. On the positive side, it facilitates the comparability of evaluation results at different times, i.e., the annotation results are comparable to and reusable in later evaluations.

Comparing ROUGE to the other evaluation methods, two of the most obvious, major advantages of ROUGE are its automatic nature and generality. In contrast to utility-based evaluation, for instance, it does not rely on information about whole sentences, but computes overlaps between texts, thereby accounting for summarization systems that are not purely extractive – a benefit shared by the Pyramid method and Factoid analysis. When contrasting ROUGE and the two latter approaches, in turn, one of the key differences is the depth of knowledge required to make similarity assessments. While ROUGE relies on n-gram overlaps between the reference summaries and the system summary, Factoid analysis and Pyramid evaluation use semantic similarity for determining like informational content. Although the automation of the former is quite straightforward as it focuses on surface realizations, the emphasis on semantic relations results in a more accurate assessment of the informational content. However, given existing methods – the Pyramid scheme and Factoid analysis in particular – the cost of this higher accuracy is considerable, because the approaches are completely manual and, as a direct implication, evaluation outside the main evaluation event is very difficult because annotations tend to differ (at times significantly) across individuals.

ROUGE-BE (Hovy et al., 2006), a version of ROUGE that relies on basic elements as individual units as opposed to the n-grams of other versions of ROUGE, strives to move beyond the focus on surface realizations towards the use of syntactic (and some semantic) information to achieve greater accuracy when evaluating peer summaries. It comprises three main components: the first creates basic elements, the second determines the similarity between two basic elements, and the third assigns scores to individual basic elements. Basic elements (BE) are either the heads of the major syntactic constituents within the given summary or a relation between a head-BE and a

single dependent. The parse trees are obtained using the Charniak parser (Charniak, 2000), Minipar (Lin, 1998), a chunker, or the Microsoft parser (Heidorn, 2000).

Comparing this approach to the one proposed in this chapter, the main difference is the dissimilar base for evaluation. While ROUGE-BE relies on simple relations (independent of the content of the other reference summaries), both the Pyramid evaluation and the system developed below are based on atomic units of information the size of which depends on the various reference summaries. The other difference is the composition of the syntactic relations being utilized.

The upshot of the discussion up to this point is that, given their general applicability and consideration of semantic content, the indicated manual evaluation methods are most accurate in their assessment of system summaries' informational content, though at the cost of significant human effort. The optimal evaluation method would therefore be an automatic version of the Pyramid evaluation scheme, owing to the more generous scope of its informational units compared to Factoid analysis.[1] A recent attempt to this end is the evaluation scheme proposed by Harnly et al. (2005), who attempt to partially automate the Pyramid evaluation scheme.

Like the present work, Harnly et al. (2005) aim to automate Step 2 of the Pyramid scheme, i.e., the matching of the SCUs from a manually constructed pyramid into system summaries. They pursue this task by using uni-gram similarity, single-link clustering, and dynamic programming to find previously determined SCUs in peer summaries. To be precise, their algorithm has the following four steps.

1. Enumerate *all* potential contributors;
2. match the most similar SCU to each contributor;
3. select from the set of candidate contributors, a covering, disjoint set of contributors that have maximum overall similarity with the Pyramid; and
4. score the system summary on the basis of the selection in the preceding step.

In terms of implementation, they depart from the annotation guidelines of the Pyramid evaluation scheme insofar as they, for computational purposes, impose a contiguity requirement – a contributor can only contain a single piece of text without breaks (as are allowed in the original Pyramid method). By imposing this constraint, they significantly reduce the number of potential contributors since not all possible combinations

[1] A second reason is the availability of more data for the Pyramid annotation scheme.

of the words in the sentence (factorial number) can be generated. Nonetheless, they show that their approach approximates the manual Pyramid evaluation method better than ROUGE.[2]

Although closely related to the present work in the sense that they consider the same task, there are a number of distinct differences between the approach by Harnly et al. (2005) and the approach developed in this chapter. For one, while Harnly et al. (2005) enumerate all possible contributors for the sentences in the summary, my approach tries to match the SCU in the sentence directly, thereby avoiding the computational explosion caused by enumeration. Likewise, to avoid the need for very similar syntactic structure between sentences, I utilize a limited set of syntactic relations, which capture the main relations between the entities and events in the sentences (only). The third key difference is that I use WordNet information to incorporate lexical semantic information into my system. Even though WordNet potentially introduces ambiguity, the use of WordNet in conjunction with syntactic relations reduces the negative influence of WordNet.

3.3 The Architecture and its Implementation

In order to found the fully automatic evaluation system pursued in Chapter 4 on a solid footing, as a first step, I construct an algorithm to determine and methodically match SCUs. To this end, the remainder of this chapter presents and ascertains the quality of a procedure to match the manually-generated SCUs from given pyramids into the associated system-generated summaries. The main focus in this regard is on detecting whether a sentence contains similar information to that within a particular SCU. Although, in some sense, this task is similar to the task of matching summary sentences to the sentences in the original document (Copeck et al., 2006), in the present case, one cannot assume a high degree of surface similarity between the surface realizations of the SCUs in the reference and system summary texts. On account of this complication, the present methodology requires use of fairly deep NLP processing techniques to determine similarity. For transparency, this section presents the architecture underlying the proposed procedure, and expends some time on its implementation. Sections 3.4 and 3.5 subsequently discuss in some detail each of the identified component steps.

[2]Their syntactically motivated preliminary experiments did not seem to lead to SCUs that are similar to the manual SCUs. In their ultimate approach, however, they utilize all syntactic relations provided by a dependency parser.

3.3.1 Architecture

The complete architecture to match SCUs from a given manually generated pyramid into a system summary is depicted in Figure 3.2. It involves three main steps: pre-processing, main processing, and evaluation. The "original documents" shown at the top of the figure contain the human reference summaries as well as the system summaries (all annotated according to the Pyramid method; for an example of a Pyramid annotation of the reference summaries, refer to Appendix B and for an example of the annotation of a peer summary with respect to a given pyramid, refer to Appendix C). As a first step, the documents are pre-processed using LT-TTT2 (Grover and Tobin, 2006), which provides sentence boundaries, tokenization, lemmatization, and named-entity information, ENJU (Sagae et al., 2007), which provides syntactic information, and WordNet (Fellbaum, 1998), which provides synset information for all nouns, verbs, adjectives and adverbs. All of the ensuing information is integrated into the overall data structure for each of the document sets; the structure ultimately contains lemmatization and WordNet information on the word level, and named entity, syntactic, pyramid annotation information as hierarchical models (pointers) rooted on the individual words, and the manual Pyramid annotation information. Figure 3.3 illustrates the information available after pre-processing for a sample sentence.

The data structure, manually constructed syntactic templates (*cf.* Section 3.5.3), and WordNet relations beyond synset information comprise the input to the main processing stage. First, syntactic templates are instantiated on the reference and peer summary. In a second step, the SCUs from the peer summary are matched into the peer summary sentences using the template instantiations and WordNet. The result of this process is a full pyramid annotation, i.e., the peer summaries now contain annotations denoting the matching of the SCUs into the sentences. The annotations can then be used to obtain a score for each annotated peer summary, which constitutes the evaluation stage of the architecture.

3.3.2 Implementation

The architecture was implemented using the object-oriented programming paradigm, Java in particular. The pre-processing classes (one for each step such that good reusability and extendability could be achieved) wrap the interaction between the data structure and the external programs. They take the original documents, convert them into the hierarchical data structure that is exemplified in Figure 3.3, and each wrapper class

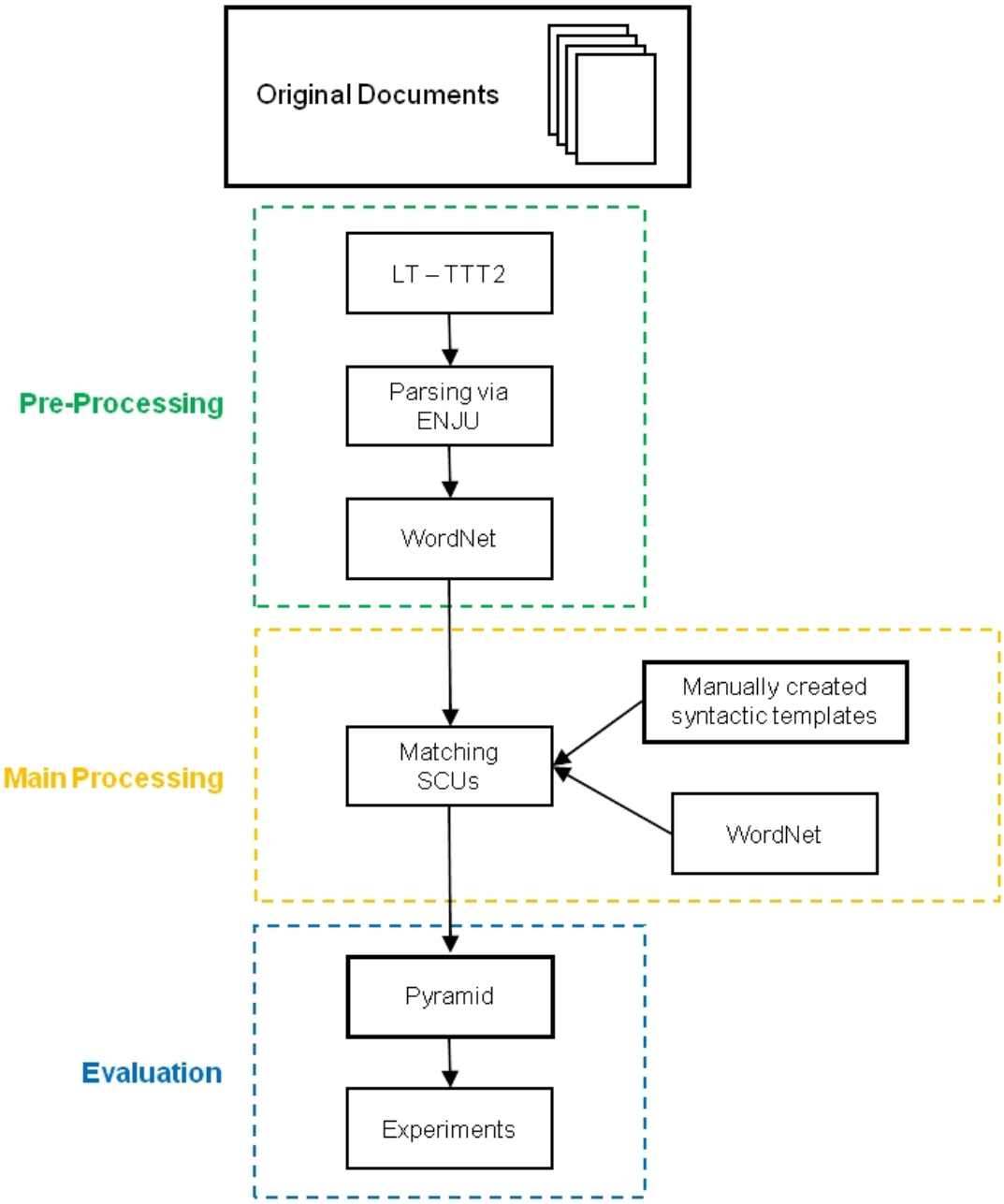

Figure 3.2: The general architecture for the SCU matching process. Boxes with bold frames represent data structures, while boxes with normal frames represent processing steps.

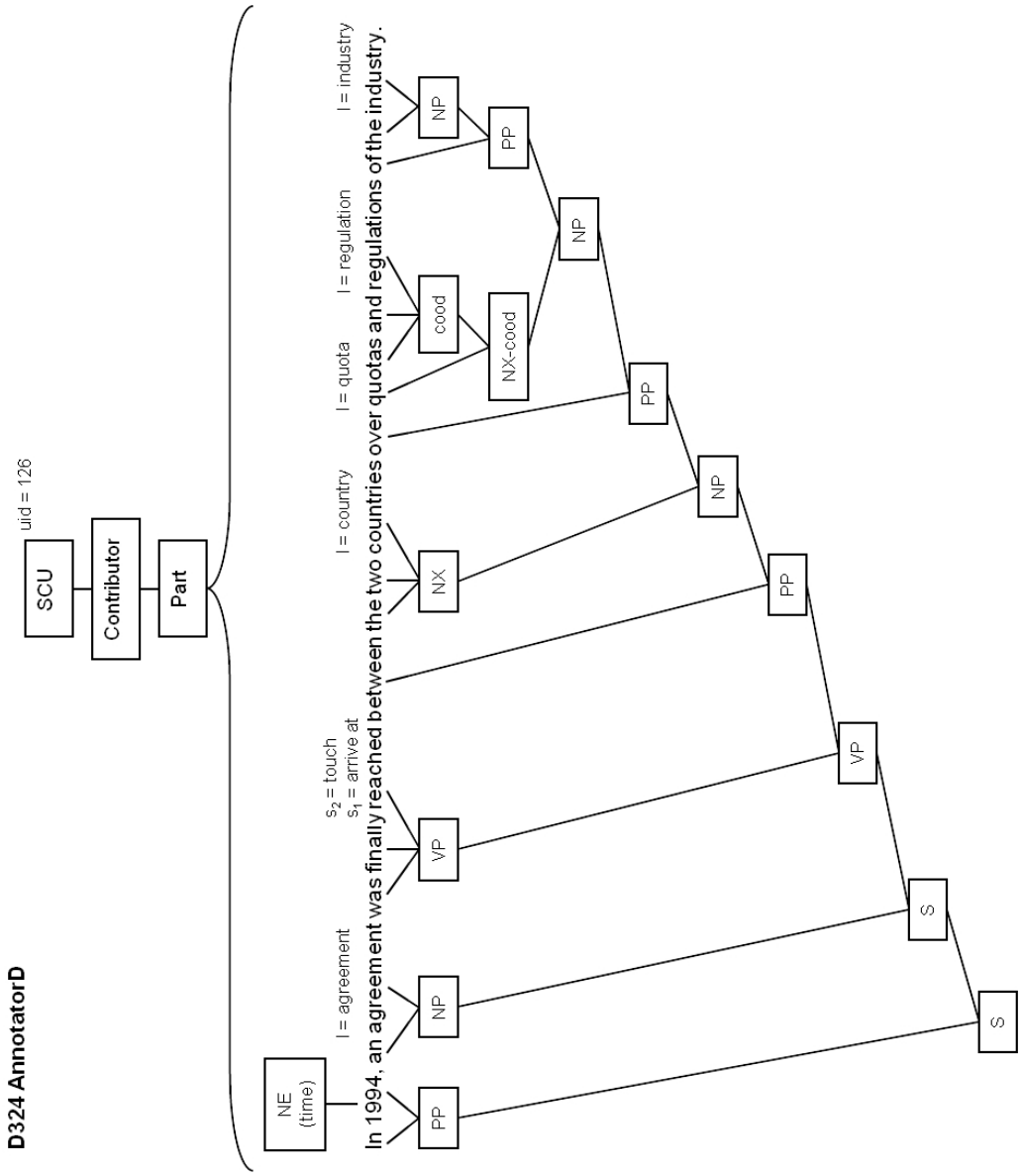

Figure 3.3: Sentence annotation following pre-processing. The sample sentence is taken from the summary of document collection D324 of DUC2005 created by human annotator D. The figure illustrates the information available following the Pre-Processing steps, including the syntactic information (below the sentence; simplified for readability), named-entity information (1994; above the sentence), lemmatization information (the 'l' attributes on the nouns; simplified for readability), the synset information (the 's' attributes on the verb; simplified for readability), and the Pyramid annotation information (above the sentence).

converts the object-oriented data structure into the required input format, starts the external processing, and integrates the result into the data structure. The main matching step ("Matching SCUs," Figure 3.2) regulates the matching process on a sentence and SCU basis, i.e., it controls how the SCUs are matched into the sentences by establishing whether the comparison is on a contributor or SCU level. Main processing also provides for the annotation of the peer summaries with the information derived from the matching process.

The most interesting aspect, however, is the encoding of the constraints on the syntactic structure of the templates. The four classes illustrated in Figure 3.5 control this aspects of the implementation. Note that they are, in principle, capable of describing a wide variety of syntactic trees and abstractions of such trees, say, by allowing variable distances between a parent and its child. The main class describing a syntactic template is the Template class. The main components of this class are its match() method, which determines the matches of the template in a particular subgraph identified by the Node parameter. The syntactic template itself is identified by three lists of conditions (represented by instances of the class Condition). The first list, "included," presents the conditions the syntactic template has to fulfill, while "excluded" defines that the conditions in that list cannot be present for the syntactic template. The third list, "conditionalExcluded," in turn, contains conditions to be ignored in order to determine a match; it is useful, among others, in order to ignore relative clauses given by noun phrases. The Condition node is the access to the actual syntactic structure encoded by way of the ConditionNode and the SyntacticConditionNode. Both encode the syntactic structure using further ConditionNodes. Note that ConditionNodes can allow for gaps in the syntactic structure in the sense that a child does not necessarily need to be defined, but instead a subgraph of the child can be defined.

Recall that the matching process relies on the use of syntactic templates to encode which syntactic transformations should be considered identical. As discussed in detail in Section 3.5.3, the templates are based on the hierarchical model used to encode relevant information such as part-of-speech and syntactic categories in the document (*cf.* Figure 3.3). As such, the syntactic templates are capable of representing any information that can be represented via the hierarchical model; Figure 3.4 provides an example of template instantiation. The most important more advanced features of the templates are the provision of generalization capabilities and exclusion filters. The generalization capabilities are applied on the word and the clause level. The former allows for variations of surface realizations such as synonyms, while the latter allows for fuzzy

matching of syntactic structures, say, by only specifying that particular sub-structures need to occur within two steps of the current node. The exclusion filters are mainly of practical importance, as I realized that full sub-clauses are often fully attached to the verb phrase, thereby allowing matching over clause boundaries. Exclusion filters provide an ability to restrict this behavior.

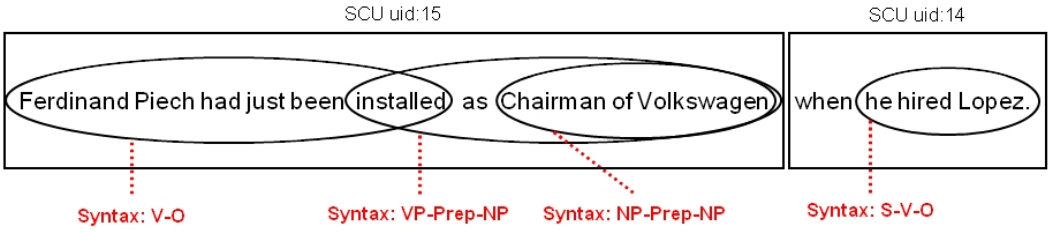

Figure 3.4: An example of template instantiations in the sample sentence “Ferdinand Piech had just been installed as Chairman of Volkswagen when he hired Lopez.”

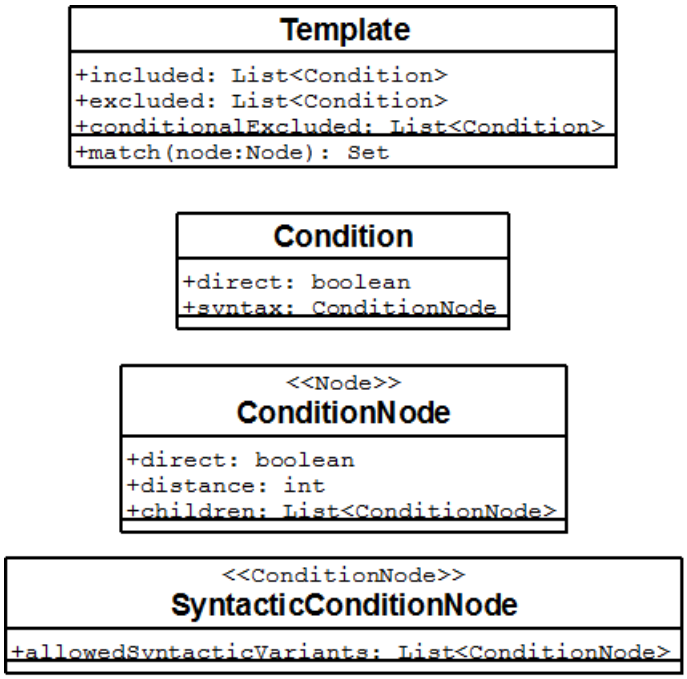

Figure 3.5: The class diagrams used to implement the proposed architecture.

3.4 Pre-Processing: LT-TTT2, Parsing, and WordNet

Given the general architecture of the proposed methodology, the remainder of this chapter provides more detail for each of the individual steps by presenting and contrasting a number of possible techniques to achieve the envisioned result. In this light,

as indicated, pre-processing consists of three straightforward steps. First of all, the original documents – source documents (where required), reference summaries, and system summaries – are processed using LT-TTT2 (Grover and Tobin, 2006), which provides tokenization, lemmatization, part-of-speech tagging, and chunking. Second, the reference and system summaries are parsed using the HPSG parser ENJU (Sagae et al., 2007) based on the sentence boundary information provided by LT-TTT2. I decided to use ENJU because an informal test indicated that it is more accurate than a number of other parsers, among others, the Collins parser (Collins, 1997), Minipar (Lin, 1998), and RASP (Briscoe et al., 2006). Furthermore, RASP (Briscoe et al., 2006) was used for the track "Recognising Textual Entailment" of the Text Analysis Conference and yielded disappointing results (Shen et al., 2008). Note, however, that a detailed comparison of the different parsers is beyond the scope of this thesis. Further advantages of ENJU are that it is based on highly accurate HPSG grammars (Miyao and Tsujii, 2008), is extremely fast, and that it outputs the results in XML format. In the third pre-processing step, all nouns and verbs are processed using WordNet (Fellbaum, 1998), thereby identifying possible synonyms, antonyms, and the like.[3]

Observe that none of these steps involves word-sense disambiguation (Yarowsky, 2000), i.e., the identification of the sense (or meaning) of a word as applicable in the context of a document. For an example of the problem, consider the word 'bass,' which has two different meanings, one being a particular species of fish and the other (vocal or instrumental) tones of low frequency. In most cases, not using word-sense disambiguation would be likely to result in numerous incorrect identifications of similarity between words. However, as words are not used in isolation but in their syntactic context, in the present case, an incorrect match between the two senses would not be critical. Since words in a syntactic context need to be similar in order to result in a match for the overall syntactic structure, the possibility of incorrectly matching these senses in their respective contexts is low. On the other hand, if word-sense disambiguation is performed and the wrong sense is selected, it might result in two occurrences of the same words in similar syntactic context not being detected. As a result, the restriction of word-sense disambiguation is actually likely to decrease the matching of syntactic structures.

[3]WordNet is a lexical database of English, which groups nouns, verbs, adjectives, and adverbs into cognitive synonyms (synsets), each expressing a distinct concept. The main relations in WordNet are: (1) synonyms are words with identical or very similar meaning; (2) antonyms are words with opposite meaning; and (3) a hyponym shares a type-of relationship with its hypernym. In WordNet, senses of a word denote different meanings of the same lexical word.

To clarify the usefulness of the second and third pre-processing steps and the information obtained by them, consider the following two examples (also used for DUC2005).

Example 1: Parsing. Figure 3.6 illustrates the benefits of using syntactic information for determining similarity of the system SCUs to the SCUs in the pyramid rather than basing the assessment on an n-gram matching model with proximity constraints. In particular, the syntactic information derived via the parsing step associates the subject of the sentence (Dr. White) and the described event (killed) in both of the sample sentences. An n-gram matching model with proximity constraints, in turn, would not recognize the relationship between 'Dr. White' and 'killed' in the second sentence because of the subordinate relative clause between the words.

Dr. White (S) killed (V) 12 people (O).
Dr. White (S), who is department head at the hospital, killed (V) 12 people (O).

Figure 3.6: Parsing. The advantage of using syntactic information as opposed to n-gram models for determining similarity of SCUs. An n-gram matching model with proximity constraints would not recognize the relationship between subject and verb in the second sentence.

Example 2: WordNet. Figure 3.7 exemplifies the effect of lexical semantic relations derived via WordNet on the identification of potential matches between informational units. It highlights that lexical semantic relations can exceed mere synonym expansion and, in fact, might improve match rejection based on antonym relations. In other words, the use of WordNet as a second pre-processing step allows for the *rejection* of some potential matches not because of low similarities but because of high dissimilarities. As such, it actually utilizes semantic relations to determine opposite meanings. In this particular example, both "killed" and "murdered" as well as "individuals" and "patients" are related to one another via WordNet. Any similarity measure that does not, in some way, condense multiple words to the same meaning (be it dimensionality reduction in latent semantic analysis or the use of WordNet) cannot hope to capture these similarities. WordNet was chosen for the present purpose as it provides manually created information in addition to a wider variety of relations between words, such as antonyms, which will be useful in the main processing steps as opposed

to dimensionality reduction methods that rely on co-occurrence statistics.

Dr. White (S) killed (V) 12 individuals (O).
Dr. White (S) murdered (V) more than 10 patients (O).

Figure 3.7: WordNet. The advantage of using WordNet information as opposed lemma or surface form information for determining the similarity of different sentences. Using WordNet, "killed" and "murdered" can be associated with each other, as can "individuals" and "patients."

3.5 Main Processing: Exploiting Linguistic and Annotation Information

The preceding section illustrated the (theoretically) positive impact of syntactic parsing and lexical semantics on the identification of similar information expressed using different surface forms. This section uses these techniques to create a method that automatically matches SCUs into sentences. In particular, given a collection of sentences, the algorithm determines whether a particular SCU contributor actually matches a particular sentence. To this end, it considers the syntactic templates along with the lexical semantics of the words in the constituents of the template, and assesses the matching of individual contributors between instantiations of the templates in order to determine whether the overall template matches.

To achieve this objective, a number of individual steps must be accomplished.

1. It is necessary to construct a method for determining *potential* matches for individual words from the SCU contributors in the pyramid and the sentences in the peer summary.

2. The information about potential matches of individual words obtained as part of the previous step, can then be used to devise an approach to identify similarity between constituents of template instantiations between pyramid and peer summary sentence.

3. The information on the constituent matches, in turn, can be exploited to develop an effective means for determining overall similarity between template instantiations between pyramid SCUs and peer summary sentence.

Pseudo-code for the overall process is provided in Figure 3.8. Besides clarifying the matching process between template instantiations, it demonstrates the complexity inherent in comparing *all* sentences and *all* SCU contributors to each other. In the following, I provide details and different options to achieve each individual step of the process.

```
use the manually created syntactic templates to convert
 SCU to templates and augment SyntacticConditionNodes
 with allowed syntactic variants based on manual
 templates, e.g, X's Y vs. Y of X

iterate over templates
    iterate over all sentences
        if (sentence is high-level (first step) match)
            iterate over potential subtrees
               if (subtree matches template)
                   annotate subtree as match to template
```

Figure 3.8: Comparison of potential subtrees to the template of a given SCU.

3.5.1 Word-Based Contributor Matching

Note that in the context of creating a highly accurate matching between the content of the SCU contributors and the SCUs in the system summaries virtually all potential methods for determining similarity in meaning are, in their essence, based on words and their individual meanings. For maximal success, several of the available methods are typically combined into more complex entities. To uncover the method(s) yielding the most accurate matching for the purposes at hand, a first step for more complex subsequent similarity measures, I experiment with a variety of word-similarity measures, including the following:

- the exact (inflected) surface forms of the words;
- the lemmas, where available, of the surface forms (derived via the lemmatization information provided by LT-TTT2);

- a variety of WordNet relations, including hypernyms, hyponyms, antonyms, and synonyms;
- content words (open class words) only; and
- words *not* included on a stop-word list.

While the first two measures are straightforward, the use of WordNet to determine similarity between words is complicated by a number of problems. Two of the most intricate issues in this regard are hypernym/hyponym derivations and word-sense disambiguation, both of which pose significant theoretical problems. As a consequence, I limit the former to direct cases and cases in which there is a direct link between two words, i.e., there are no sibling relations along the path between the words,[4] and omit the latter (*cf.* Section 3.4). The actual algorithm for determining a match based on WordNet relations is show in Figure 3.9.

The measures based on content words and a stop-word list, in turn, are essentially self-explanatory. Content words are words in the open part-of-speech classes, i.e., nouns, verbs, adjectives, and adverbs, whereas a stop-word list is a list of frequently used words that, in and of themselves, do not contribute to the meaning of a sentence. The criteria for these two measures often tend to overlap insofar as stop-word lists contain all closed part-of-speech class words (e.g., pronouns and prepositions) as well as frequent open-class words such as the verb 'to be' or the noun 'thing.'

The results of this step are useful in two ways. For one, they allow one to narrow down the number of sentences that potentially contain a specific SCU such that more in-depth processing can be performed on specific peer summary sentences and SCU pairs. Second, the information can be used in the matching of constituents of syntactic templates, which constitutes one of the core parts of the proposed matching algorithm.

3.5.2 Constituent Matching

The objective of the next phase is to determine sets of syntactic templates and approaches to match corresponding constituents between templates. The ideal scenario for this undertaking would be two completely separate steps. In practice, however, a

[4]For example, for inquiry [a search for knowledge], the two hyponyms 'experiment' [the testing of an idea] and 'investigation' [an inquiry into unfamiliar or questionable activities] would not result in the identification of a match between the two hyponyms. However, between each of the individual hyponyms and the hypernym, a match would be identified.

```
//X represents the actual word
wordX: the surface form of word X
sensesX: the different word senses of wordX
treeX: the word senses in the path from wordX
    to its highest parent node
hypernymsX: the direct hypernyms of wordX
hyponymsX: the direct hyponyms of wordX

match = false
//find match if one of the word senses are synonyms
for (i = 0; i < senses1.length; i++)
  if (senses2.contains(senses1[i])
    match = true
//find match if one word sense is the direct hypernym of
 another
for (i = 0; i < senses1.length; i++)
  if (hypernyms2.contains(senses1[i])
    match = true
//find match if one word sense is the sibling of another
for (i = 0; i < hypernyms1.length; i++)
  if (hypernyms2.contains(hypernyms1[i])
    match = true
//find match if one word is in the path of the other
if (tree1.length > tree2.length)
  longer = tree1
  base = senses2
else
  longer = tree2
  base = senses1
for (i = 0; i < longer.length; i++)
  if (base.contains(longer[i]))
    match = true
```

Figure 3.9: The algorithm for determining word-similarity using WordNet.

strict separation is hardly feasible as the size and composition of the constituents depends intrinsically on the specific syntactic templates used as a basis for the constituent matching. Nonetheless, conceptually, one can distinguish a number of algorithms for determining the compatibility of two SCU constituents. (Note that pronouns are resolved to potential antecedents. Potential antecedents are considered to be those noun phrases that occur in the preceding sentence or the same sentence before the pronoun. The subsequent analysis is then carried out for all potential matches. In future work, full anaphora resolution might be performed, *cf.* Chapter 7.) Those most appropriate for the purpose at hand include:

- the determination of the percentage of relevant words (be it lemmas, surface forms, or WordNet derivations) from the SCU contributor's constituent that are also found in the potential Peer-SCU's constituent;
- the assessment of the compatibility of the constituents' head words (derived via the semantic head attributes provided by ENJU) using WordNet;
- the assessment of the compatibility of the constituents' head words (derived via the semantic head attributes provided by ENJU) using WordNet when excluding conflicting modifiers, i.e., modifiers that imply opposite meaning (antonyms);
- the constituents' deconstruction to determine smaller sized constituents – in the case of appositions, for instance, each of the noun phrases are smaller constituents.

3.5.2.1 Percentage of Relevant Words

The method of matching constituents based on the percentage of relevant words is quite straightforward. As the name suggests, it determines the similarity of two constituents on the basis of the percentage of words in the constituent from the SCU contributor that are also contained in the constituent from the peer summary. If the percentage exceeds a manually determined threshold level δ, the algorithm concludes that the two constituents are equivalent.

The algorithm for the matching process is demonstrated in Figure 3.10. The threshold level δ, i.e., the constituents are considered equivalent if 65% of their content overlaps, was set to 0.65 based on the performance of the algorithm during an exploration of the threshold values in the interval [0.5,1] in steps of 0.05. Note that matching

based on both contributors is relevant in situations in which the contributors have very different lengths.

3.5.2.2 Compatibility of Head Words

More often than not, syntactic templates are based on the syntactic relations of various noun and verb phrases, both of which contain head words, i.e., words that are particularly relevant to convey the meaning of the phrase. As such, a possibility for determining the similarity of two constituents is to assess the compatibility of the constituents' head words. The necessary comparisons can readily be achieved by way of the semantic head annotation provided by ENJU during pre-processing. However, the exclusive consideration of head words may result in important conflicting information in the modifiers of the head word being ignored (for example, "the 80-year-old designer" and "the 70-year-old designer" both share the same head, yet, the modifiers provide a strong indication that the two designers are not the same individuals unless the frame of reference for both expressions is different – one references a time period ten years earlier). The strictest way to ensure that modifiers are not conflicting is the incorporation of antonym relations from WordNet. Figure 3.11 summarizes the relevant algorithm. It is quite straightforward in that it matches the heads as identified by ENJU and, if they match, performs an antonym detection in an optional step.

3.5.2.3 Deconstruction of Constituents

While there are good reasons for using either of the preceding methods for matching constituents, both also have notable (potential) problems. When basing one's assessment on head words, for instance, phrases such as 'a lot of effort' and 'a group of Germans' would have "lot" and "group" as head words. Clearly, however, the important attributes of the sample sentences are "effort" and "Germans." Along similar lines, the assessment of the percentage overlap can be misleading if the constituents of the SCU are long and contain additional information.

An approach not as exposed to these problems is the deconstruction of a constituent into its building blocks. Since the deconstruction process not only considers the individual noun and verb phrases within a constituent individually, but also considers the whole constituent to be a match if one of the individual noun or verb phrases matches, the matching of the individual noun and verb phrases is again based on the percentage and head-word compatibility methods. As indicated in Figure 3.12, which provides

```
matches = 0
total1 = number of relevant words in the first
 contributor
total2 = number of relevant words in the second
 contributor

for all relevant words in the first contributor
  for all relevant words in the second contributor
    if current words in both contributors match
      matches = matches + 1
percentage = matches / total1

matches = 0
for all relevant words in the second contributor
  for all relevant words in the first contributor
    if current words in both contributors match
      matches = matches + 1

if percentage < (matches / total2)
  percentage = matches / total2

if (percentage > δ)
  return true
```

Figure 3.10: Constituent Matching based on the Percentage of Relevant Words.

```
head1 = head word of first constituent
head2 = head word of second constituent

if head1 matches head2
  if NOANTONYMDETECTION
    return true
  else
    for relevant words in first contributor
      for relevant words in second contributor
        if current words are antonyms
          return false
    return true
return false
```

Figure 3.11: Constituent Matching based on the Compatibility of Head Words.

pseudo-code for the relevant algorithm, the matching of the individual noun phrases is achieved using the percentage overlap measure (δ) described above, as it performed slightly better.

3.5.3 Syntactic Templates

Up to this point, the focus was largely on word-level and chunk information of head words. One's understanding of a text, however, patently requires more than merely a collection of noun and verb phrases. Rather, the relation between these chunks of words is critical.

Harnly et al. (2005) have, in this regard, indicated that using a full parse-tree output does not help in the task of emulating the Pyramid evaluation method. Likewise, Lin (2004) achieves similar results when using Rouge-BE, which exploits syntactic information, as he does with his Rouge-N and/or Rouge-SU-N methods, both of which do not involve syntactic information. It therefore seems that syntactic information in contexts of their research is not particularly helpful. As such, the use of a full parse-tree appears to obscure the main important information, i.e., the main relations between chunks.

Therefore, I limit the syntactic information available to the evaluation method and define a number of simple templates that capture specific syntactic phenomena and

```
nps1 = noun phrases of first constituent (maximum noun
 phrases only containing only words or noun phrases in
 the parse tree)
nps2 = noun phrases of second constituent

for all nps in nps1
  for all nps in nps2
    if the current nps match
      return true
return false
```

Figure 3.12: Constituent Matching based on the Deconstruction of Constituents.

collapse the syntactic information for the individual constituents (i.e., the parts of the templates that are filled in the instantiation) such that no syntactic information within a constituent is utilized (Figure 3.13). In particular, I consider the following (small) set of basic syntactic templates to be used in varying combinations to identify the template(s) that provide the most accuracy without matching too much irrelevant information:

Basic Syntactic Templates

- **SVO:** Subject – Verb – Object
- **SV:** Subject – Verb
- **VO:** Verb – Object
- **NN:** Noun Group, Noun Group (Apposition)[5]
- **NPrepN:** Noun Group – Preposition – Noun Group[6]
- **XPrepX:** Any Group – Preposition – Any Group

The first three templates are obvious and, in fact, exceedingly similar. The reason for considering all three as separate entities is that they allow for a number of different possibilities for capturing objects. That is, a different number of objects can be captured by using multiple Verb – Object templates, one for each object. Appositions,

[5]In future experiments, this template is not explicitly used, but incorporated into the deconstruction of constituents.

[6]This template also covers possessives and multi-word expression nouns.

besides constituting potential replacements for Subject – Verb – Object constructs, also facilitate the determination of alternative expressions for the same entity. The fifth template, Noun Group – Preposition – Noun Group, among others, accounts for nominalizations. Finally, Any Group – Preposition – Any Group, also (amongst other things) captures a variety of realizations of objects, e.g., "Piech was elected as chairman of VW," where 'as' fills the preposition slot, while 'elected' and 'chairman of VW' fill the 'Any Group' slots, resulting in the correct identification of the relation between the election event and the post to which he was elected.

In addition to constructing the templates themselves, it is necessary to specify a way to determine similarities between instantiations of the templates. In other words, it is necessary to determine which constituents of one template correspond to constituents in another template. Without these correspondences, there is no way to determine similarities between *different* templates. For the purposes at hand, the constituents are correlated using manually created information, i.e., it is determined manually which constituents in one template, if any, corresponds to which constituents in another template. For example, an annotator establishes that the subjects and verbs, respectively, correlate between the first two templates, while Verb – Object and Any Group – Preposition – Any Group correlate 'Verb' with the first 'Any Group' and 'Object' with the second.

3.5.4 Information-Sharing between Different SCU Contributors

The final, theoretically promising, avenue investigated for the purposes of matching SCUs between an existing pyramid and the associated peer summary is the extent of information shared between different contributors. In particular, I exploit different realizations of a given SCU in order to create a more comprehensive representation of it, thereby enabling the matching process to include information from different contributors (to obtain a match). In principle, the approach entails combining the information from all contributors in an SCU into a single, coherent data structure – I call it "concept." A descriptive diagram of a concept is given in Figure 3.14. As shown, a concept consists of multiple instantiations of syntactic templates along with groups of constituents that correspond to the same underlying entity or event.

Although similar in spirit, this approach takes a slightly different outlook on the task to be completed than the methods described so far. The aforementioned methods compare two instantiations of syntactic templates directly, i.e., one instantiation from

```
st1 = instantiation of syntactic template
st2 = instantiation of syntactic template

mapping = get the potential mappings between the
 constituents in the two syntactic templates (from
 manually created file)

for all the possible mappings in mapping
  matches = true
  for all constituents in st1
    if the corresponding constituent (from mapping) in
     st2 does not match
      matches = false
  if matches
    return true
  matches = true
  for all constituents in st2
    if the corresponding constituent (from mapping) in
     st1 does not match
      matches = false
  if matches
    return true
return false
```

Figure 3.13: Matching Syntactic Templates.

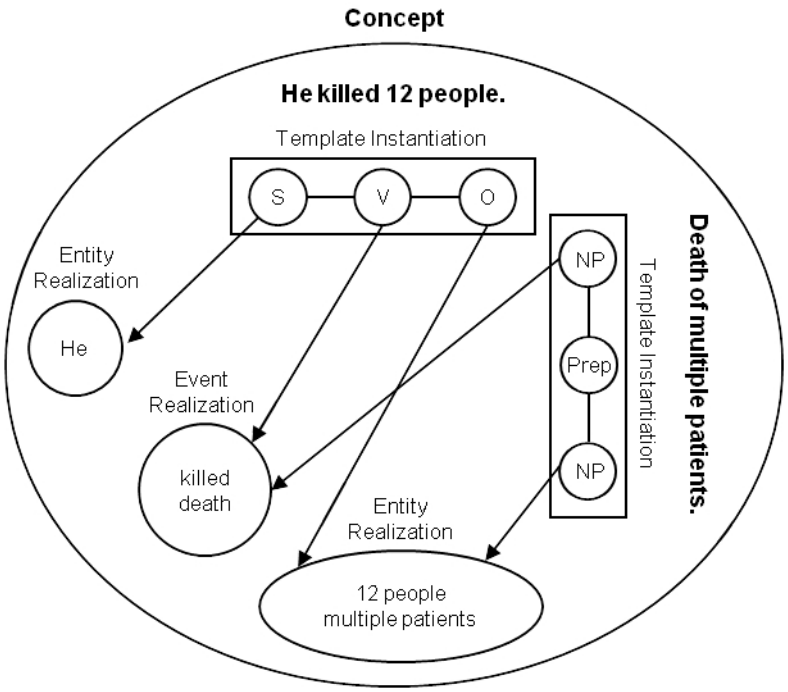

Figure 3.14: The ”Concept.” The relationship between underlying concept, syntactic realization, and entity/event realization.

an SCU contributor with an instantiation from the peer summary. In the information-sharing setting, the instantiations of the different SCU contributors are compared to each other and, where appropriate, combined into concepts. A given instantiation from a peer summary is then compared to these concepts. For illustration, consider the following two SCU contributors: “He killed 12 people” and “Death of multiple patients.” The appropriate concept would be the one depicted Figure 3.14. Now, if the peer summary contained “He killed multiple patients,” it is not hard to see that the use of the concept would result in a better overall similarity between the peer summary and the concept than if the peer summary unit was compared against the individual SCU contributors. Figure 3.15 provides a synopsis of the matching algorithm involving information sharing using pseudo-code. Note that it considers alternative surface realizations for each of the constituents.

To sum up, this section explored a number of different methods to determine similarity between informational units, starting with simple word-based approaches and incrementally building a more comprehensive picture of the informational content of pieces of text into a single data structure called ‘concept.’ The following section investigates the practical performance of the various measures in a number of experiments. They initially assess the performance of the indicated measures by themselves. Subse-

```
concept = a concept that is to be matched
concept.syntactic = list of syntactic templates
concept.syntactic.constituents = ordered list of the constituents
 of the syntactic template
concept.syntactic.constituents.realizations = list of realizations
 of the constituent in different template instantiations

st = instantiation of syntactic template to be matched

for all syntactic realizations in the template
  mapping = get the potential mappings between the constituents in
   the current syntactic template of the concept and st (from
   manually created file)
  for all the possible mappings in mapping
    matches = true
    for all constituents in st1
      aMatch = false
      for all corresponding realizations (from mapping) in
       concept.syntactic.constituents.realizations
        if corresponding constituents matches
          aMatch = true
      if not aMatch
        matches = false
    if matches
      return true

    matches = true
    for all constituents in concept.constituents
      aMatch = false
      for all realizations in concept.constituents.realizations
        if the corresponding constituent (from mapping) in st
         matches
          aMatch = true
      if not aMatch
        matches = false
    if matches
      return true
return false
```

Figure 3.15: Matching Concepts.

quently, the correlation of the ranking of summaries based on the combination of the best measures is compared to other manual and automatic evaluation approaches.

3.6 Evaluation: Experiments

In general, the evaluation of a system-generated summary results in it being assigned a score capturing its quality relative to a number of human reference summaries. When summaries from competing summarization systems are evaluated, the scores can be used to create a ranking of the systems, with the best system producing the highest scoring summaries. The investigation of summarization evaluation methods typically uses these rankings to correlate the ranking produced by a given evaluation method with other established evaluation methods as well as a manual gold standard. Paralleling the scores assigned to system summaries, the assessment of evaluation methods results in a ranking of the methods, with the most accurate method assigned the highest score.

In view of its consistently reliable results, the gold standard used for comparison in the present setting is the results of the evaluation of the summaries using the manual Pyramid evaluation method (Nenkova and Passanneau, 2004). As the objective of the various algorithms to be tested here is the partial automation of the Pyramid scheme *given* a manually-created pyramid, one has a number of other viable avenues for evaluation. For one, one can assess the precision and recall (Olson and Delen, 2008) of matching SCUs into sentences. Expanding on this, one can investigate the accuracy with which the human SCU contributor boundaries correspond to the system contributor boundaries, i.e., not only determining whether a sentence contains an SCU, but establishing that a particular set of words represents the SCU contributor. My evaluation is based on the precision, recall, and ranking correlation as the primary evaluation metrics; I do not pursue the finer-grained possibility, which is appropriate in the present case because the focal point is whether the information is present (per se) as opposed to where it is in a given sentence.

3.6.1 The Datasets and Experimental Procedure

The usefulness of various combinations of the algorithmic approaches described in previous sections for the purposes at hand was evaluated using the multi-document summarization datasets from DUC2005 (Dang, 2005). They consist of 50 clusters

containing a total of 300 human summaries. Each cluster, in turn, comprises documents pertaining to a single set of events, e.g., the corporate espionage case between Volkswagen (VW) and General Motors (GM) in the 1990s. I used a subset of 10 clusters as a development set for the investigated linguistic phenomena and their relative importance, leaving the remaining 40 clusters for the techniques' evaluation (referred to as test set below).

I evaluate the proposed system by way of five specially designed experiments. The first two are dedicated to exploring the precision and recall on a word-based level as well as a contributor-based level, respectively. In line with the general definition, precision is defined as $\frac{TP}{TP+FP}$ and recall as $\frac{TP}{TP+TN}$, where TP denotes the true positives, FP the false positives, and TN the true negatives with respect to the identification of the SCUs (e.g., Yakushiji et al., 2005). Note that precision only requires one of the contributors to be matched in the sentence for a match to occur, while recall considers each contributor individually. The third experiment investigates the usefulness of sharing information between different contributors of the same SCU, again, using precision and recall as evaluation metrics. The fourth investigates the correlation between the ranks of the system-generated summaries in the overall ranking of the summaries using different combinations of the aforementioned methods and the ranks according to the official Pyramid and ROUGE evaluation schemes. The final experiment explores the performance of the proposed system in the context of a formal evaluation task and dataset. Namely, the AESOP – "automatically evaluating summaries of peers" – dataset originally introduced to the Text Analysis Conferences (TAC) in 2009, described in detail in Section 3.6.6. The main objective of this investigation is the comparison of the proposed approach to other automatic evaluation methodologies.

3.6.2 Experiment 1: Word-Based Contributor Matching

This experiment, which is based on the 10-cluster development set, investigates the percentage of words from the SCU contributors that can be matched with the sentences in the peer summaries. The underlying (practical) idea is that in order to accurately determine whether an SCU is present in a sentence, it is a necessary requirement that a significant portion of the words in the SCU are matched into the sentences containing that particular SCU. For that reason, the percentage is more important than the precision of the matching. Over-matching will be addressed by way of the syntactic templates investigated in the subsequent experiments.

The results for this experiment are summarized in Table 3.1. For each measure and/or combination of measures, it reports two scores: (1) the percentage of words in the SCU contributors that can be matched in the peer summary SCUs (or more precisely, the sentence containing the SCU; called PeerSCUs), and (2) the percentage of the words in the PeerSCUs that can be matched in the SCU contributors. Note that "perc" denotes the global percentage, while "avg-perc" designates the average of the percentages for the individual pairs. In general, the results indicate that the use of content words only and stop-word lists tends to improve performance, the combination of both yielding better results than each individually. The notable exception is the WordNet case, where the combination of content words and stop-word lists does not improve performance compared to using content words only.

Method	**avg-perc$_{\text{SCU}}$**	**perc$_{\text{SCU}}$**	**avg-perc$_{\text{Peer}}$**	**perc$_{\text{Peer}}$**
surface	0.36	0.32	0.45	0.36
surface + stop	0.40	0.40	0.51	0.41
surface + content	0.39	0.36	0.52	0.41
surface + content + stop	0.43	0.42	0.54	0.44
lemma	0.37	0.32	0.47	0.36
lemma + stop	0.41	0.40	0.53	0.40
lemma + content	0.42	0.38	0.54	0.43
lemma + content + stop	0.45	0.43	0.57	0.44
WordNet	**0.54**	**0.50**	0.47	0.37
WordNet + stop	0.46	0.41	0.55	0.39
WordNet + content	**0.54**	**0.50**	**0.62**	**0.49**
WordNet + content + stop	0.53	0.48	0.60	0.45

Table 3.1: Results of Experiment 1. The percentage overlap between SCU contributor and peer summary SCUs (called PeerSCU). $perc_{SCU}$ denotes the percentage of the words in the SCU contributor that are also in the PeerSCU, while $perc_{PeerSCU}$ represents the percentage of the words in the PeerSCU that are also in the SCU contributor. "surface" denotes surface form, "lemma" canonical form, "content" content word, and "stop" stop-word list.

The results provide strong support for the conjecture that the use of additional knowledge resource, WordNet in particular, has benefits for recall. To be precise, using WordNet increases the system's recognition rate by more than 5%. Correspondingly,

the experiments in the following use the "WordNet (synonym/hypernym/hyponym) + content" setting for word recognition.

3.6.3 Experiment 2: Syntactic Templates and Constituent Matching

As discussed, the next stage in the automatic matching of SCU contributors in peer summaries is based on various sets of syntactic templates and approaches to matching corresponding constituents. Since these two aspects cannot be separated because templates determine the size and composition of the constituents, I investigate a number of combinations of these measures. The selected combinations along with the results on the same subset of 10 document clusters as used for Experiment 1 are displayed in Table 3.2. The first column indicates the template(s) used in the particular combination of measures, the second column the way in which constituent matching is performed (*cf.* Section 3.5.2), and the third column the method used to determine whether two words are sufficiently similar in a semantic sense.

The results show that head word and percentage matching appear to give rise to very similar results, while the deconstruction of the constituents is by and large marginally better. In addition to this observation, the use of the full set of templates yields the best performance. This seems to indicate that the development of more templates may result in further improvement of performance. However, it will be important not to enumerate all possible templates since this would essentially defeat the purpose of the templates, which is to limit the number of syntactic structures that are used in the evaluation.

3.6.4 Experiment 3: Information-Sharing between Different SCU Contributors

Based on the findings regarding the template composition and matching, this experiment combines information from different contributors into one single comprehensive data structure that shares information between the different templates and contributors. It involves two parts, both exploring the same issue, but the first using the 10-cluster development dataset and the second using the 40-cluster test dataset. Table 3.3 summarizes the results when using the development dataset. It shows that using the proposed conceptual representation for information sharing results in improved recall for the

Templates	**Constituent**	**Word**	**Percentage Matching**
SV	percentage	Exact	0.75
SV	percentage	Lemma	0.78
SV	percentage	WordNet	0.80
SV	head word	WordNet	0.78
SV	deconstruct	WordNet	0.81
VO	percentage	Exact	0.76
VO	percentage	Lemma	0.76
VO	percentage	WordNet	0.78
VO	percentage	WordNet	0.78
VO	head word	WordNet	0.77
VO	deconstruct	WordNet	0.78
SVO	percentage	Lemma	0.81
SVO	head word	WordNet	0.81
SVO	deconstruct	WordNet	0.82
NPrepN	percentage	WordNet	0.75
NPrepN	head word	WordNet	0.74
NPrepN	deconstruct	WordNet	0.74
XPrepX	percentage	WordNet	0.76
XPrepX	head word	WordNet	0.75
XPrepX	deconstruct	WordNet	0.75
SV VO SVO	percentage	WordNet	0.85
SV VO SVO	head word	WordNet	0.85
SV VO SVO	deconstruct	WordNet	0.86
SV VO SVO NPrepN	percentage	WordNet	0.90
SV VO SVO NPrepN	head word	WordNet	0.91
SV VO SVO NPrepN	deconstruct	WordNet	0.90
SV VO SVO NPrepN XPrepX	percentage	WordNet	0.92
SV VO SVO NPrepN XPrepX	head word	WordNet	0.92
SV VO SVO NPrepN XPrepX	deconstruct	WordNet	**0.93**

Table 3.2: Results of Experiment 2. The percentage of matches for the correct identification of PeerSCUs. In the third column, “Exact” denotes the use of the surface forms and “Lemma” the use of the base forms of the surface realizations.

identification of PeerSCUs.

The majority of PeerSCUs not correctly identified are extremely short PeerSCUs that do not contain any templates. An example of a relevant SCU would be 'GM's,' a PeerSCU indicating the fact that Lopez was formerly an employee of GM. The full sentence underlying this example is, "The issue stems from the alleged recruitment of GM's eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues." Hence, given only the information in the contributor, it is impossible to determine this employee relationship because the possessive might refer to any number of things, ranging from employee to plant to announcement (in the context of the automotive industry). One could also argue that the possessive 'GM's' by itself does not accurately represent the informational content of the SCU and as such represents a spurious annotation that requires significant thought regarding the boundaries of the SCU contributor.

Method	**Recall**	**Precision**
Identification *with* Information Sharing	0.95	0.90
Identification *without* Information Sharing	0.93	0.90

Table 3.3: Results of Experiment 3 (Part 1). The impact of information sharing on the precision and recall of the PeerSCU identification process using the best method identified in the preceding experiments on the development dataset. Precision and recall are computed at the SCU-level, i.e., there is a positive identification if *any* of the contributors of the SCU is matched in the relevant sentence.

At this point, we have a method for identifying PeerSCUs that achieves a high recall. However, it is important that high recall is achieved without sacrificing precision. Even though the syntactic motivation should guard against this problem, Table 3.3 also provides insight into this question. Namely, the results indicate that information sharing does not result in degraded precision. On the contrary, the underlying matching method shows good precision for the identification of PeerSCUs. The remainder of the errors is mostly caused by short SCU contributors that result in overmatching the particular facts.

Table 3.4 explores the issue of information sharing using the 40-cluster test dataset. It shows that the experiments in this chapter have resulted in a matching method for SCUs that is highly precise while at the same time providing good recall on unseen data. The performance on the test dataset in comparison to the performance on the

Method	Recall	Precision
Identification *with* Information Sharing	0.92	0.90
Identification *without* Information Sharing	0.91	0.88

Table 3.4: Results of Experiment 3 (Part 2). Precision and recall for the detection of PeerSCU contributors on the test dataset.

training dataset (Table 3.3) also shows that the performance loss on unseen data is relatively small. This indicates that the method should perform well in general as opposed to a small subset on which it was specifically developed. In addition, it further supports the conclusion that information sharing improves recall of SCU information.

3.6.5 Experiment 4: Performance of the Proposed Methodology

The penultimate experiment investigates the overall impact of the methods investigated in the previous experiments on the detection of PeerSCU contributors. To this end, I determine the ranking correlation between the semi-automatic method developed in this chapter and the original, manual Pyramid method. This experiment, too, is based on the 40-cluster test set of the DUC2005 dataset that are annotated using the Pyramid evaluation method. The results are summarized in Table 3.5. They show that the results of my method constructed using the preceding preliminary experiments translate well into the actual detection of PeerSCU contributors. In fact, the rank correlations show that my method outperforms both ROUGE and the purely word-based hierarchical clustering approach by Harnly et al. (2005).[7]

3.6.6 Experiment 5: Evaluation Using AESOP2009 Dataset

The foregoing experiments developed and evaluated the proposed semi-automatic evaluation system based on DUC2005 Pyramid annotation information. As the DUC2005 set was originally released for the purpose of creating summarization systems, it had to be manipulated manually to suit the evaluation scenario, implying that comparability with other systems is quite limited. In order to establish the system's performance relative to as many evaluation systems as possible, this experiment is based on the

[7]There is a minor problem with the comparison in Table 3.5 in that the result for Harnly et al. (2005) is computed on the whole 50-cluster DUC2005 dataset, while the other results are computed on a 40-cluster subset of the DUC2005 dataset.

Method	Ranking Correlation
Identification *with* Information Sharing	0.97
Identification *without* Information Sharing	0.96
ROUGE-2	0.93
Word-Based Clustering (Harnly et al., 2005)	0.95

Table 3.5: Results of Experiment 4. (Spearman ρ) Ranking correlation between my semi-automatic methods, ROUGE, and Harnly et al. (2005)'s word-based clustering approach with respect to the original manual Pyramid method on the test dataset. The issue being investigated is the similarity of the rankings obtained by each of these methods compared to the rankings obtained using the manual Pyramid method.

purpose-built AESOP2009 dataset, which allows for a comparison with (up to) thirty-nine competing systems.

In principle, the DUC2005 and AESOP2009 datasets are very similar in structure. Both consist of topic statements, a number of document sets, and several human-authored reference (or model) summaries. In particular, the AESOP2009 dataset consists of a total of 44 topics, each of which comprises two document sets containing ten documents each, and four reference summaries. In view of the specific task motivating its release – the automatic assessment of automatic multi-document summarization systems – the AESOP2009 dataset, moreover, comprises a number of automatically generated summaries created as part the TAC2009 "Update Summarization" task. As part of this task, automatic multi-document summarization systems were to summarize the first document set and, based on the assumption that the first document set/summary is known, use the second document set to create an "update" summary. Note that since the *semi*-automatic evaluation method introduced and developed in this chapter requires manually created pyramids as input, I add the manually generated pyramids developed for evaluation of the update summarization systems to the resources of the evaluation dataset.

The task set to assess the performance of automatic evaluation systems is the production of two summary-level scores:

1. a score for assessing the quality of the informational content of all of the peer *and* reference summaries, referred to as "AllPeers," the task being designed to reveal the systems' ability to distinguish between the human and automatic summaries;

and

2. a score for each of the peer summaries only, referred to as "NoModels."

As in the preceding experiment, the evaluation is carried out by computing the similarity between the rankings obtained by way of the automated evaluation methods and the ranking obtained via the manual Pyramid evaluation scheme. In this case, however, a total of three correlation metrics are computed – namely, the Pearson, Spearman, and Kendall correlation coefficients – which are formally defined as follows (for two rankings r_a and r_b):

***Pearson r*:**

$$r(r_a, r_b) = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $\bar{X}$ represents the mean of variable X and X_i denotes the i^{th} value of variable X.

***Spearman* ρ:**

$$\rho(r_a, r_b) = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}$$

where d_i signifies the difference in ranks of corresponding values and n stands for the total number of values in the dataset.

***Kendall* τ:**

$$\tau(r_a, r_b) = \frac{P - Q}{P + Q}$$

where P represents the number of concordant pairs (i.e., the number of pairs ordered in the same way in both rankings) and Q denotes the number of discordant pairs (i.e., the number of pairs ordered differently by the two rankings).

In general, the approaches by the participants of the AESOP2009 task vary considerably. One set of the approaches, for instance, learns regression models over different ROUGE scores (Conroy et al., 2009), while another determines the different concepts in the model summaries using a taxonomy and concept similarity metrics and then determines the overlap to the concepts in the system summaries (Steinberger et al., 2009). Other approaches utilize character n-grams and overlap or edit-distances between the n-gram representations for different documents (Giannakopoulos and Karkaletsis, 2009), statistical distributions of words in the different documents (Kumar and Kumar, 2009), or generative models using a variety of different features (Katragadda, 2009). One of the approaches also considers the sentence as the atomic unit and subsequently determines whether the peer summary sentences are sufficiently

similar to any of the sentences in the model summaries to assess the quality of the summary (Ouyang and Li, 2009).

The evaluation approach that is most similar to the approach presented in this thesis is an extension to ROUGE-BE, called ROUGE-BEwT-E (Tratz and Hovy, 2009). This approach differs from ROUGE-BE in two ways: (1) it considers the presence of a basic element in the model summaries, but only assigns binary weights (a basic element either occurs in the model summaries (1) or does not (0)), and (2) a number of transformations are used to determine similarity between two basic elements. The major differences to the approaches presented in this thesis are that the basic elements are not of variable size (i.e., basic elements are not combined to create elements with more content), that the elements are not weighted, and that the transformations and syntactic structures considered are different. As regards the latter, my systems consider adjectives and adverbs to be of minor importance and none of the templates include adjectives or adverbs as constituents, although they might be important in (and are considered as part of) the matching of the constituents.

Given this background, the ranking correlations of my partially automated evaluation system along with those of twenty-four (other) fully automatic evaluation methods are summarized in Tables 3.6, 3.7, 3.8 and 3.9. In particular, the tables show the correlation between the ranking according to the automated evaluation approaches and that according to the manual Pyramid methods for both scores – i.e., AllPeers and NoModel – for the initial and update summaries. The results are sorted according to Kendall's τ metric (in Column 4).[8] An inspection of the various tables reveals that no single approach consistently achieves the highest correlation with the ranking according to the manual Pyramid method. While the results for the scenarios based on the initial summaries AllPeers as well as NoModels are very similar, those for the evaluation of the update summaries differ distinctly from those for the initial summaries.

The ranking correlations obtained via my partially automated evaluation approach for the initial *and* update summaries are consistently good. The only competing systems to provide uniformly good results are the ones with ID 12 and 15, though, the consistency achieved by my approach is slightly better. Considering that the other approaches are fully automatic, it will be interesting to investigate the performance of my fully automated Pyramid-style approach (*cf.* Chapter 4). In terms of the Pearson and Spearman correlation coefficients, a number of systems achieve very high results, i.e.,

[8]The sorting is according to Kendall's τ as it does not make assumptions regarding the distributions of the variables and as the scores do not imply that the correlations are close to perfect.

System ID	**Pearson (r)**	**Spearman (ρ)**	**Kendall (τ)**
Partial Automation	0.980	0.957	**0.851**
12	0.969	**0.962**	0.847
6	0.706	**0.962**	0.835
11	0.982	0.953	0.826
15	0.581	0.954	0.823
26	0.797	0.953	0.821
13	0.781	0.948	0.814
4	0.802	0.947	0.813
10	0.640	0.947	0.812
1 ROUGE-SU4 (baseline)	0.734	0.946	0.811
28	0.618	0.950	0.809
17	**0.983**	0.936	0.800
18	0.884	0.936	0.797
24	0.978	0.933	0.796
31	0.966	0.938	0.795
19	0.885	0.932	0.789
5	0.558	0.938	0.789
25	0.628	0.935	0.786
16	0.884	0.929	0.783
33	0.468	0.926	0.780
2 ROUGE-BE (baseline)	0.586	0.919	0.779
23	0.772	0.925	0.778
34	0.877	0.913	0.766
27	0.826	0.907	0.748
8	0.743	0.900	0.738

Table 3.6: Results of Experiment 5 (Part 1). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for AllPeers when using the initial summaries. The results are sorted according to Kendall's τ metric.

System ID	**Pearson (r)**	**Spearman (ρ)**	**Kendall (τ)**
28	0.656	**0.966**	**0.858**
Partial Automation	0.901	0.940	0.857
15	0.614	0.961	0.847
16	0.856	0.942	0.821
19	0.858	0.940	0.818
24	**0.978**	0.941	0.817
12	0.967	0.944	0.811
11	0.976	0.942	0.807
18	0.841	0.932	0.805
23	0.739	0.937	0.802
10	0.647	0.940	0.801
26	0.792	0.929	0.798
2 ROUGE-BE (baseline)	0.629	0.934	0.796
25	0.665	0.934	0.794
6	0.704	0.931	0.792
5	0.610	0.919	0.781
31	0.963	0.920	0.777
33	0.534	0.914	0.770
13	0.756	0.905	0.757
1 ROUGE-SU4 (baseline)	0.726	0.901	0.754
17	0.973	0.896	0.753
8	0.740	0.892	0.749
4	0.752	0.893	0.745
27	0.799	0.895	0.743
22	0.951	0.900	0.741

Table 3.7: Results of Experiment 5 (Part 2). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for AllPeers when using the update summaries. The results are sorted according to Kendall's τ metric.

System ID	**Pearson (r)**	**Spearman (ρ)**	**Kendall (τ)**
Partial Automation	0.910	0.949	**0.835**
6	0.911	**0.950**	0.820
12	0.901	0.947	0.815
26	**0.978**	0.942	0.810
15	0.805	0.939	0.800
11	0.954	0.933	0.796
10	0.869	0.931	0.793
2 ROUGE-BE (baseline)	0.857	0.936	0.791
4	0.967	0.928	0.788
25	0.850	0.928	0.787
1 ROUGE-SU4 (baseline)	0.921	0.923	0.785
13	0.952	0.924	0.785
28	0.830	0.933	0.785
18	0.965	0.918	0.770
19	0.967	0.917	0.769
23	0.928	0.908	0.765
17	0.952	0.908	0.759
31	0.894	0.912	0.758
5	0.799	0.915	0.757
32	0.815	0.901	0.755
33	0.742	0.900	0.752
24	0.963	0.902	0.750
16	0.962	0.900	0.742
21	0.796	0.887	0.730
34	0.897	0.873	0.715

Table 3.8: Results of Experiment 5 (Part 3). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for NoModels when using the initial summaries. The results are sorted according to Kendall's τ metric.

System ID	**Pearson (*r*)**	**Spearman (ρ)**	**Kendall (τ)**
Partial Automation	0.890	0.920	**0.843**
28	0.908	**0.955**	0.841
15	0.887	0.950	0.831
2 ROUGE-BE (baseline)	0.924	0.932	0.801
25	0.896	0.937	0.800
16	0.968	0.918	0.789
10	0.918	0.924	0.785
12	0.946	0.923	0.781
24	0.957	0.916	0.781
19	0.962	0.911	0.781
23	0.932	0.912	0.776
5	0.895	0.920	0.774
11	**0.970**	0.918	0.772
18	0.944	0.901	0.768
26	**0.970**	0.903	0.768
33	0.855	0.919	0.768
13	0.962	0.904	0.754
6	0.921	0.902	0.753
8	0.937	0.880	0.734
31	0.940	0.884	0.734
4	0.946	0.874	0.719
1 ROUGE-SU4 (baseline)	0.940	0.863	0.708
21	0.652	0.848	0.702
17	0.944	0.847	0.698
27	0.934	0.854	0.695

Table 3.9: Results of Experiment 5 (Part 4). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for NoModels when using the update summaries. The results are sorted according to Kendall's τ metric.

results approaching 1, while the results for Kendall's tau are far below perfect agreement. The differences between the performances according to the different metrics are a result of the metrics' peculiarities. The most problematic is the Pearson correlation, since it provides very different results compared to the other two metrics. This is likely caused by the assumption of normally distributed data.

3.7 Sample Passages Highlighting Strengths and Weaknesses of my Evaluation System

In their essence, the experiments presented in the previous section explored the performance of the partially automated Pyramid-style evaluation system developed in this chapter from an aggregate perspective. They showed that the system compares well to other (fully automated) evaluation systems. The objective of this section is to analyze the merits and weaknesses of the system by way of a number of representative sample document passages. All of the sample sentences are taken from DUC2005 documents. They are selected to illustrate situations in which the system works well, as well as scenarios in which the system does not correctly recognize similarity, or in which the content units' similarity is beyond the scope of the patterns recognized by the system.

The first representative portion of text, presented in Figure 3.16, illustrates a number of issues within the system's scope of ability as well as some of the shortcomings. Each of the figures containing an example is organized as follows: The top presents the manually identified SCU that is relevant to the setting, then the manually identified PeerSCU that is related to the SCU at the top. Lastly, at the bottom is the summary sentence in which the PeerSCU was identified, i.e., the sentence in which the partially automated approach should identify the SCU.

On a general note, the last two contributors of the SCU – "drugs" and "narcotics" – reveal that the Pyramid annotation of the present passage is problematic, since both only contain a single word. Nonetheless, Contributor 5, "and have been trained to detect narcotics," is a good example to illustrate the matching process underlying the system. In the original sentence, the subject is "dog," thus providing both the Subject – Verb and Verb – Object relations required for a successful match. The Subject – Verb match is problematic because of the difficulty to recognize "detect" and "smell out" as similar in meaning, which can only be determined using the definition of the synset. The Verb – Object relation, in turn, can only be successfully matched because of the

```
<scu uid="146" label="Dogs are used to sniff out narcotics">
 <contributor label="to sniff narcotics">
  <part label="to sniff narcotics" start="443" end="461"/>
 </contributor>
 <contributor label="Dogs trained to sniff out narcotics">
  <part label="Dogs trained to sniff out narcotics" start="2388" end="2423"/>
 </contributor>
 <contributor label="their use in detecting narcotics">
  <part label="their use in detecting narcotics" start="4955" end="4987"/>
 </contributor>
 <contributor label="to sniff out evidence of drugs">
  <part label="to sniff out evidence of drugs" start="6629" end="6659"/>
 </contributor>
 <contributor label="and have been trained to detect narcotics">
  <part label="and have been trained to detect narcotics" start="10461" end="10502"/>
 </contributor>
 <contributor label="drugs">
  <part label="drugs" start="8576" end="8581"/>
 </contributor>
 <contributor label="narcotics">
  <part label="narcotics" start="3562" end="3571"/>
 </contributor>
</scu>

<peerscu uid="146" label="(7) Dogs are used to sniff out narcotics">
 <contributor label="Drug-sniffing dogs">
  <part label="Drug-sniffing dogs" start="0" end="18"/>
 </contributor>
</peerscu>

Sentence in which the PeerSCU is identified:
Drug-sniffing dogs at Dover and other international borders will have to work harder
 in a frontier-free Europe as illegal traders look for a boom in single market crime.
```

Figure 3.16: Example 1. Representative portion of document 112.D426.M.250.A.1. (DUC2005) featuring manual Pyramid annotation

use of WordNet, which determines that "drug" is a direct hypernym of "narcotic." Hence, there is a Subject – Verb – Object pattern match between the fifth contributor and the sentence containing the PeerSCU. The best match for the SCU and PeerSCU, however, is provided by the second contributor ("Dogs trained to sniff out narcotics"), which only requires the use of WordNet to determine similarity between the objects.

Attempting to match against the fourth contributor ("sniff out evidence of drugs") illustrates one of the as yet unresolved problems of the matching process, because the Noun – Prep – Noun relation currently does not detect a similarity of the verbs' object relations because of the prepositional attachment of the "drugs". Accordingly, for this contributor, none of the suggested templates provide for a matching of the contributor and PeerSCU. The problem is compounded by the use of different words for the informational unit, as well as the different syntactic structure. The approach could probably cope with substantial differences in one of the two areas, but both exceed the capabilities of the approach.

The second example, shown in Figure 3.17, illustrates the problems associated with named entities. In the PeerSCU, the robotic aid is explicitly referred to as "Handy 1," and the only reference to the topic of robotics is in the project's name. Without the use of information sharing, the best shot at identifying the match between the SCU and sentence is when using the first contributor ("A robotic arm was developed to enable severely disabled people to feed themselves"). Between the contributor and the sentence, two Verb – Object relations centering on "enable" and one Subject – Verb relation centering on "eat." Can be identified that are used to match the sentence and SCU. Using information sharing, a link between "help," "allow," and "enable" can be identified, which might be useful in the later matching to other sample sentences; these links can be identified based on the syntactic structure of the contributors and their respective labels.

The contributors in the SCU in the third example, provided in Figure 3.18, are generally very similar in the expression of the same information as shown by their very similar wording and syntactic structure. The most similar contributor is the second contributor, "Until recent times, Malaysia was the world's largest producer of tin." In it, the Subject – Verb – Object relation centering on "was," a Noun – Poss – Noun relation centering on the possessive, and a Noun – Prep – Noun centering on "of" can be identified. "Until recent times," on the other hand, does not participate in any relation because its first argument is the full sentence as opposed to a noun group as required by the template definition. All of the relations except the Noun – Prep – Noun

```
<scu uid="28" label="Robotic arms are used to help disabled people feed themselves">
 <contributor label="A robotic arm was developed to enable severely disabled people
  to feed themselves">
  <part label="A robotic arm was developed to enable severely disabled people to
   feed themselves" start="1459" end="1540"/>
 </contributor>
 <contributor label="Robots have been made that can feed disabled patients">
  <part label="Robots have been made that can feed disabled patients" start="4263"
   end="4316"/>
 </contributor>
 <contributor label="Robots can feed and provide other personal assistance to the
  disabled">
  <part label="Robots can feed and provide other personal assistance to the
   disabled" start="9607" end="9676"/>
 </contributor>
 <contributor label="include robotic arms that allow a disabled person to feed
  himself">
  <part label="include robotic arms that allow a disabled person to feed himself"
   start="11408" end="11473"/>
 </contributor>
 <contributor label="and a robotic arm with...has been developed for the disabled
  who are unable to feed themselves">
  <part label="and a robotic arm with" start="8098" end="8120"/>
  <part label="has been developed for the disabled who are unable to feed
   themselves" start="8139" end="8208"/>
 </contributor>
 <contributor label="Severely disabled persons eat more comfortably when fed by a
  robotic">
  <part label="Severely disabled persons eat more comfortably when fed by a robotic"
   start="3025" end="3093"/>
 </contributor>
</scu>

<peerscu uid="28" label="(6) Robotic arms are used to help disabled people feed
 themselves">
 <contributor label="Handy 1, designed by Mike Topping, development manager
  University of Keele's rehabilitation robotics project, enables severely disabled
  people to eat unaided">
  <part label="Handy 1, designed by Mike Topping, development manager University of
   Keele's rehabilitation robotics project, enables severely disabled people to eat
   unaided" start="252" end="409"/>
 </contributor>
</peerscu>

Sentence in which the PeerSCU is identified:
Handy 1, designed by Mike Topping, development manager University of Keele's
 rehabilitation robotics project, enables severely disabled people to eat unaided.
```

Figure 3.17: Example 2. Representative portion of document 113.D431.M.250.H.10. (DUC2005) featuring manual Pyramid annotation

relation can be identified in the sentence containing the PeerSCU and thus result in a successful match.

The sample passage presented Figure 3.19 again illustrates a problem with named entities that needed to be addressed in the development of the system – the use of abbreviations. In the first contributor as well as the peer sentence, both "Salvation Army" and "SA" are identified as named entities. However, the proposed system needs to identify the abbreviation as similar/identical in order to match them correctly. A similar situation arises for the detection of partial names; in the example, "Zahn Memorial Center for Social Services" versus "Zahn." The text fragment also illustrates that the system developed in this chapter correctly identifies passive constructs. What is more, this example demonstrates that a correct identification is not only possible if the SCU has many contributors, but also if the contributors comprise low-weight SCUs.

Overall, the examples explored in this section provided an overview of the working of the partially automated Pyramid-style evaluation scheme developed in this chapter. They also illustrated a number of issues that have not been explicitly addressed in the development section, e.g., the identification of similarities between named entities, and highlighted the merits of information sharing, which facilitates the development of links between a number of different verbs for which no links in WordNet could be identified. Based on the results of the experiments, the problems identified in the examples in this section did not cause a detrimental failure of the evaluation approach. However, if these and related issues are addressed in future work, the overall performance of the evaluation approach should increase further.

3.8 Discussion

The algorithm presented in this chapter constituted a first step towards the automation of the Pyramid evaluation scheme for the informational content of automatically generated summaries. The results show that the matching of the SCUs into the document by way of my method performs very well. In particular, I have shown that the ability to generalize on the word level using lexical semantic knowledge such as synonyms and hypernyms provides for increased similarities between human and peer summaries. The overgeneralization of using WordNet and not using word-sense disambiguation can be balanced by using syntactic templates. The false overgeneralization on the word level only results in incorrect template instantiation matches if all constituents are incorrectly overgeneralized. Last but not least, I introduced a construct for infor-

```
<scu uid="252" label="Malaysia used to be the world's premier producer of tin">
 <contributor label="Until recently, it was the world's premier producer">
  <part label="Until recently, it was the world's premier producer" start="901"
   end="952"/>
 </contributor>
 <contributor label="Until recent times, Malaysia was the world's largest producer
  of tin">
  <part label="Until recent times, Malaysia was the world's largest producer of tin"
   start="1641" end="1709"/>
 </contributor>
 <contributor label="Malaysia became the world's leading tin producer">
  <part label="Malaysia became the world's leading tin producer" start="3601"
   end="3649"/>
 </contributor>
 <contributor label="Malaysia was once the world's leading tin producer">
  <part label="Malaysia was once the world's leading tin producer" start="4704"
   end="4754"/>
 </contributor>
 <contributor label="Malaysia gradually became the leading producer of tin in the
  world">
  <part label="Malaysia gradually became the leading producer of tin in the world"
   start="6350" end="6416"/>
 </contributor>
 <contributor label="Since 1857 South East Asia's Malaysia had been the world's
  largest tin producer">
  <part label="Since 1857 South East Asia's Malaysia had been the world's largest
   tin producer" start="7922" end="8001"/>
 </contributor>
</scu>

<peerscu uid="252" label="(6) Malaysia used to be the world's premier producer of tin">
 <contributor label="Malaysia was once the world leading tin producer">
  <part label="Malaysia was once the world leading tin producer" start="1017"
   end="1065"/>
 </contributor>
</peerscu>

Sentence in which the PeerSCU is identified:
Malaysia was once the world leading tin producer.
```

Figure 3.18: Example 3. Representative portion of document 115.D632.M.250.I.15. (DUC2005) featuring manual Pyramid annotation

```
<scu uid="80" label="They operate the Zahn Memorial Center for Social Services in Los
 Angeles">
 <contributor label="The SA operates Zahn Memorial Center for Social Services">
  <part label="The SA operates Zahn Memorial Center for Social Services"
   start="9189" end="9245"/>
 </contributor>
 <contributor label="They operate the Zahn Memorial Center for Social Services">
  <part label="They operate the Zahn Memorial Center for Social Services"
   start="4225" end="4282"/>
 </contributor>
</scu>

<peerscu uid="80" label="(2) They operate the Zahn Memorial Center for Social Services
 in Los Angeles">
 <contributor label="Zahn is operated by the Salvation Army">
  <part label="Zahn is operated by the Salvation Army" start="0" end="38"/>
 </contributor>
</peerscu>

Sentence in which the PeerSCU is identified:
Zahn is operated by the Salvation Army and the money will be earmarked for the L.A.
 homeless.
```

Figure 3.19: Example 4. Representative portion of document 118.D671.M.250.G.24. (DUC2005) featuring manual Pyramid annotation

mation sharing that combines the information from all individual contributors of an SCU into one general structure that contains information about which constituents of the template instantiations are considered identical between the contributors.

My results indicate that my semi-automatic method outperforms other state-of-the-art, but fully automatic, methods with regard to the similarity of the results to the original manual Pyramid evaluation scheme. A natural implication of this finding is that there is a notable benefit to using syntactic information along with lexical semantics in the evaluation of quality of the informational content of peer summaries. At the same time, however, the use of more complex automatic evaluation systems might lead to the development of systems that exploit the shortcomings of my evaluation system. In the present case, the limited use of syntactic information might be exploited by preferring textual units for the summary that conform to the syntactic structures represented by the templates to a higher degree, thereby improving possibilities for matches based on the templates. It remains to be seen how complex and successful the exploitation of this theoretical shortcoming will be.

Despite this shortcoming, performance should not drop below that of other (partially) automated approaches since the syntactic templates and over-generalization should approximately cancel each other out, resulting in a measure that is based on the similarity of the content words represented in WordNet. For this reason, I expect this measure to be superior to other current automated approaches.

Chapter 4

Full Automation of the Pyramid Evaluation Method

4.1 Introduction

Having automated the matching component of the Pyramid evaluation method by constructing an algorithm to predict the presence of a surface realization in a system summary given a manually generated pyramid, the objective of this chapter is the scheme's full automation. In particular, it endeavors to resolve the following two issues, respectively.

1. Is it, with any degree of robustness, possible to generate a pyramid automatically?

2. Can the algorithm for automatically generating pyramids be combined with the methodology for matching pyramid SCUs in a peer summary to yield a highly accurate, fully automatic evaluation measure?

The first query in essence seeks an automated version of the first step of the two-tiered Pyramid evaluation scheme (*cf.* Figure 2.14 in Chapter 2), while the second seeks a fully automatic evaluation method for system generated summaries. The benefits of both objectives should be obvious. Not only would the manual effort for the evaluation of automatic summarization systems be reduced enormously (the only manual effort remaining would be the creation of (human) reference summaries), but the scheme would retain a semantic outlook as opposed to the purely word-based automatic evaluation measures currently available. The main advantage of capturing semantic relationships between words in a document is that the ensuing procedure takes into account whether two texts are similar in terms of the meaning(s) they convey rather than whether various sets of words occur in both texts. The upshot is the assurance that the system summary contains the essence of the source documents not only in regard to the most important text fragments but also vis-à-vis their substance.

Recall that the approach to automate the second stage of the Pyramid evaluation method involved a methodology to determine whether two (syntactic) templates are similar, i.e., whether they contain similar informational content. Using these templates, the approach subsequently established whether a larger informational unit (called summary content unit, or SCU) is present in a peer summary. To this end, the approach determined a threshold regarding the percentage of template instantiations from the pyramid SCU that are present in a summary sentence – if more than the threshold percentage of template instantiations from the SCU are found in the summary sentence,

then the summary sentence is annotated as containing that SCU. In this chapter, I build on this idea in order to determine SCUs automatically.

The procedure to automate the pyramid creation step, broadly speaking, is two-tiered. First, similar information in syntactic constructs needs to be grouped together. This step uses syntactic templates and lexical semantic information, as employed in the partial automation procedure, to extract small syntactically motivated informational units. Units that are similar according to some similarity metric (to be determined during the course of this chapter) are then grouped together using pair-wise hierarchical clustering. As part of this clustering process, the individual template instantiations in the same clusters are combined using the conceptual framework described in the preceding chapter (*cf.* Figure 3.14). At this stage, the individual clusters represent simple syntactic relations between entities. However, in order to account for the variable size of the units of information that constitute the SCUs, individual clusters that frequently occur in vicinity of each other are assumed to belong to the same unit of information. Thus, the second step of the proposed approach uses the co-occurrence statistics between the clusters (with reference to occurrence in the same sentence) in order to create the ultimate pyramid of variable-size units of information.

The remainder of this chapter is organized as follows. Section 4.2 provides a survey of the usual structure of clustering algorithms, their applications in natural language processing, and outlines the most common techniques to evaluate their performance. Section 4.3 subsequently presents the proposed algorithm to fully automate the Pyramid evaluation scheme. Section 4.4 seeks to uncover the best approaches to putting the theoretical framework into practice and evaluates the resulting pyramid both as an isolated entity and in conjunction with the matching methodology developed in Chapter 3, i.e., as a fully automated evaluation method. Section 4.6 illustrates the workings of the methodology on human reference summaries from the TAC 2009 AESOP dataset. Section 4.7 concludes with a brief discussion of the main results of this chapter.

4.2 Related Work

Clustering is a machine-learning technique concerned with detecting structure in a collection of unlabeled (i.e., not categorized or annotated) data. To be precise, clustering algorithms attempt to organize data into groups that comprise members considered similar in some (pre-determined) respect. Plainly, depending on the underlying (or subsequent) purpose, the groupings can be achieved in a number of ways. While in

the case of exclusive clustering each data object can belong to at most one cluster, overlapping clustering allows individual data objects to belong to multiple clusters. Probabilistic clustering, on the other hand, assigns probabilities for the membership of data objects to individual classes, often called "soft" clustering, as opposed to "hard" clustering in the other cases.

One of the most common ways to distinguish between different clustering algorithms is to examine the method(s) used to infer the clusters from the source data (Berkhin, 2006). That is, this classification scheme is based on the process(es) used to create the clusters, the major distinctions being hierarchical clustering, partitioning relocation clustering, density-based partitioning, and grid-based partitioning. As the system developed below involves (exclusive) hierarchical clustering, this section explores the advantages and disadvantages of common approaches to clustering using this classification scheme.

4.2.1 Succinct Survey of Common Clustering Algorithms

4.2.1.1 Hierarchical Clustering

Hierarchical clustering, as the name suggests, creates a hierarchy of data objects that are closest to one another in the sense that each cluster node contains child clusters, and sibling clusters are partitioned in such a way that all points are covered by a common parent. The two most widespread approaches to generating hierarchical clusters are agglomerative and divisive clustering. Whereas the former starts with data object clusters containing a single data object each and recursively merges two or more appropriate clusters to yield a suitable hierarchy, divisive clustering commences with a single cluster containing all data objects and recursively splits the most appropriate (i.e., similar) data objects into smaller clusters. Two major advantages of hierarchical clustering are its flexibility with respect to the granularity of clustering and the ease of utilizing any form of distance or similarity between clusters. The major disadvantages, in turn, are the vagueness of the termination criterion and the greedy nature of the algorithm, which does not revisit assignments once they are made.

In contrast to general machine-learning data representations that represent the individual data objects using a number of properties called "features," hierarchical clustering algorithms often use a matrix of distances or similarities between the data objects/clusters, called a "connectivity matrix." In large-scale applications, one of the major drawbacks of hierarchical clustering is the memory requirements of the matrix. As a

result, numerous optimizations of the matrix to reduce memory needs and/or computational complexity are available (e.g., Olson, 1995).

The connectivity matrix is computed using a so-called "linkage metric" (Murtagh, 1985). The foremost linkage metrics are single link, average link, and complete link. They use the minimum, average, and maximum distance between the data objects in various clusters, respectively, to generate the connectivity matrix.[1] However, neither of these linkage metrics is the most appropriate for the present objectives. In their place, I use conceptual clustering, as part of which, a concept description is generated for each individual cluster. The concept description is a representation of the *whole* cluster as opposed to the properties of *individual* data objects as used by the other linkage metrics. This allows for the complex combination of multiple surface realizations of content into one general representation upon which the clusters are computed. While the description of the clusters in conceptual clustering can be arbitrarily complicated, in the present case, the "concepts" idea developed in Chapter 3 is used for these purposes to limit the complexity of syntactic differences and differences within the individual constituents.

4.2.1.2 Partitioning Relocation Clustering

Partitioning relocation clustering is based on the principle of dividing a dataset into several subsets and adjusting the cluster locations until a stable convergence of the cluster composition is achieved. Due to computational restrictions, these approaches commonly use iterative optimization algorithms, i.e., cluster quality is gradually improved with each iteration of the optimization process as specified by the (respective) criteria underlying the optimization. One approach, probabilistic modeling, assumes that data derives from a mixture of several populations for which distributions and priors need to be determined. An example of probabilistic clustering is the two-step iterative expectation-maximization algorithm (Dempster et al., 1977), which alternates between the expectation step (E) and the maximization step (M). The former computes the expectation of the log-likelihood using the current estimates of the latent variables, and the latter computes the parameters maximizing the expected log-likelihood based on the estimates derived via the expectation step. The estimates are then used as inputs for the expectation step in the next round of iteration.

Alternative approaches such as k-medoids and k-means clustering start with an ex-

[1] For a detailed discussion of different generic approaches to constructing connectivity matrices, refer to Fisher and Pazzani (1991).

plicit objective function. They initially select k random data objects and assign all data objects to one of k initial clusters. Given this assignment, the cluster mean or median is recomputed. These cluster measures represent the clusters for the next iteration of the optimization. The process is repeated until the cluster composition converges to a stable assignment.

In general, partitioning relocation clustering has a multitude of advantages. For example, it facilitates the encoding of complex structure. For the objectives at hand, however, the major disadvantages are the fact that soft assignments are counter-productive (because of the hard assignment of text units to specific SCUs) and that there is no apparent feature vector of independent attributes. Owing to the use of information sharing in the first step to automate the generation of the Pyramid method, the concepts change with the progressive assignment of data objects to the various clusters. Correspondingly, the relevant techniques do not (sufficiently) support the encoding of complex correlations between data objects, as is the case in the data underlying even simple pyramids. In contrast to hierarchical clustering, partitioning relocation clustering methods do not follow the processes that human annotators follow in order to construct the pyramid, i.e., they do not use direct comparisons between data objects.

4.2.1.3 Density and Grid-Based Partitioning

Density-based partitioning is rooted in the idea that "an open set in the Euclidean space can be divided into a set of its connected components" (Berkhin, 2006). According to this approach, a cluster constitutes a connected dense component that grows in the direction that density leads. As such, the resulting clusters can be of arbitrary shape. The main disadvantage of this approach is the requirement of a metric space, which disqualifies it for my purposes because information sharing between individual data objects was shown to be beneficial for the partial automation of the Pyramid scheme.

The notion underlying grid-based partitioning is similar in spirit to density-based partitioning. Yet, rather than using individual data objects, the space is initially partitioned into a grid, whereupon the individual cells of the grid are clustered. Unfortunately, this approach suffers from the same drawback as density-based partitioning and is thus also disqualified for the present work.

4.2.1.4 Discussion: Clustering Algorithms

While all of the approaches to clustering presented in this section have advantages, for two main reasons, the methodology proposed to automate the creation of pyramids uses (exclusive) agglomerative, hierarchical conceptual clustering. First, it enables one to observe directly the accuracy of individual similarity computations. Second, and most important, conceptual clustering allows for the combination of information from a number of different data objects, i.e., information sharing between templates. As stated, I, again, make use of the relevant information at this stage of the automation of the Pyramid evaluation scheme. The main disadvantage of the other clustering approaches is the assumption that individual data objects are completely independent from each other. Although they can, in principle, optimize the clustering process using this assumption, they cannot (for the same reason) incorporate the information sharing aspect.

4.2.2 Applications of Clustering in the Field of Natural Language Processing

As clustering algorithms are not (yet) commonly used in automatic summarization, let me briefly place their involvement in the proposed methodology into a broader context. Clustering has successfully been applied in a number of other areas of NLP. One such example is POS induction, the unsupervised learning of classes of words. Clark (2003) uses a partitioning relocation approach – a variation of the k-means algorithm – to be able to extract and subsequently exploit distributional and morphological information about the words in the source text(s). Clustering has also successfully been applied to the problem of word-sense disambiguation (Shin and Choi, 2004). To this end, too, the authors use k-means clustering in order to infer the sense of a word based on its collocations. In a similar spirit, Baldewein et al. (2004) use an Expectation-Maximization (EM) based clustering approach to generalize over possible fillers, with the objective of labeling semantic roles.[2] Klein (2005), in turn, employs EM-based clustering as part of his approach to the unsupervised induction of grammatical structure, combining clustering with parameter search.

Even though not pervasive, clustering has been exploited in a few recent contributions, though usually as a step in the generation as opposed to the evaluation of sum-

[2]Semantic role labeling is the task of detecting the semantic arguments associated with the predicate of a sentence and the classification of the arguments into their specific roles (Gildea and Jurafsky, 2002).

maries. One of the earliest approaches in this context is MEAD (Radev et al., 2000), which extends single-document summarization to the multi-document summarization framework using tf.idf. In particular, they represent each document using the tf.idf values above a certain threshold for each document, clustering the documents based on their similarity to the centroids. The subsequent central hypothesis for sentence extraction is that sentences containing words from the centroids are most important.

In a similar spirit, SimFinder (Hatzivassiloglou et al., 2001) involves word, named-entity, and co-occurrences of words within specified word-windows as features for clustering paragraphs. More specifically, it employs a non-hierarchical, partitioning relocation approach called exchange method (Spath, 1985), which uses a hill-climbing approach for the optimization of clustering. The clustering results subsequently constitute the input for various sentence extraction/generation methods. Marcu and Gerber (2001) cluster elementary discourse units in their multi-document summarization system by means of a C-Link clustering algorithm (Defays, 1977) involving the cosine overlap of the discourse units. Subsequently, clusters are ranked according to importance and a representative discourse unit from the most important clusters is selected to be included in the peer summary. More recently, Wang et al. (2008) use symmetric matrix factorization to cluster sentences. To construct the similarity matrix, sentence-to-sentence similarities are computed via semantic analysis – semantic role parsing, to be precise. They then use symmetric matrix factorization to group the sentences, which can be shown to be equivalent to kernel k-means.

The approach most closely related to the present work, to date the only application of clustering to the (automatic) evaluation of summarization systems, is that by Harnly et al. (2005). They propose an automation of the Pyramid evaluation scheme based on single-link clustering. However, as discussed in some detail in Chapter 3, they do not consider syntactic or semantic similarities, and rather limit themselves to words as the basis for their clustering approach.

4.2.3 Evaluation of Clustering Algorithms

Paralleling complete summarization systems, it is crucial to evaluate the quality and/or performance of clustering algorithms, potentially even more so, since their output frequently forms an input to other processing stages. However, more often than not, the evaluation of clustering algorithms is extremely difficult, amongst other things, because of issues relating to the cluster size affecting the impact of incorrect classi-

fications on the selected evaluation metric. The purpose of this section is to survey (briefly) recent research into the evaluation of clustering methods and outline the particular metrics used in this chapter.

In the context of multi-document summarization, sentence clustering is typically evaluated indirectly via the quality of the resulting peer summaries. Although this process allows for inferences regarding the influence of the chosen clustering algorithm(s) on the ultimate summaries, it does not in any way evaluate the clustering algorithm itself. One of the main problems in this regard is that indirect evaluation, in most cases, merely requires similar partial orderings in terms of the cardinality of the clusters. That is, the clusters represented in a peer summary are generally the clusters that have the most data objects.

A more fundamental problem in the context of sentence clustering is that evaluation datasets are not readily available, nor does there (as yet) exist an automatic scheme, entailing that the manual effort required to evaluate relevant clustering algorithms is enormous. Geiss (2009) recently attempted to create a gold-standard sentence-clustering dataset. In the present context, however, their gold standard is not useful, for two main reasons: (1) their approach is based on the assumption that the appropriate unit of clustering is (necessarily) a sentence, and (2) any given sentence can only belong to exactly one cluster. In contrast, the Pyramid evaluation method (Nenkova and Passanneau, 2004) partitions the content units of human reference summaries into clusters (SCUs) containing variable-sized units of content (based on their similarity to the content of the reference summaries). In most cases, this entails that sentences are not the appropriate unit of measurement because sentences contain more than one SCU, which contradicts both assumptions in Geiss (2009).

The main problem associated with variable-sized units of content when it comes to evaluating clustering algorithms is that it is not possible to evaluate directly both unit size and clustering quality. In order to overcome this problem, I do not evaluate on SCUs directly, but instead break SCUs into smaller syntactic units (using templates), such as Subject – Verb – Object or Noun Group – Preposition – Noun Group, which describe relations between individual units, but are fixed in size. The result is a stable basis for evaluating the quality of my clustering algorithm regardless of the size of the units.

The basis for evaluation is illustrated in Figure 4.1. For the sample sentence, four units for clustering are determined based on a parse of the sentence using ENJU (Sagae et al., 2007) – in other words, the syntactic templates developed in the preceding chap-

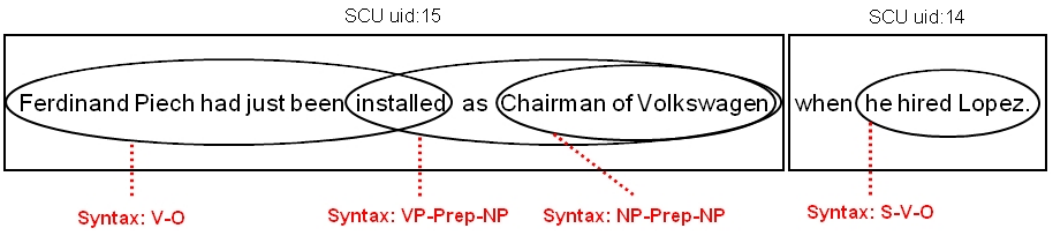

Figure 4.1: Clustering Quality. The basis for the evaluation of clustering quality is formed by simple syntactic structures, as illustrated using the sample sentence: "Ferdinand Piech had just been installed as Chairman of Volkswagen when he hired Lopez."

ter are applied to the sample sentence. Four instantiations of the templates have been determined, as illustrated by the ovals with red, bold labels. Three of these units are in the SCU with ID 15 (black box on the left-hand side), while the fourth unit is in the SCU with ID 14 (black box on the right-hand side). The quality of clustering of the template instantiations is captured by the overlaps between the gold standard – determined using the Pyramid information and the template instantiations within the SCUs – and the clusters determined by the clustering algorithm. Paralleling other approaches, the main vulnerability of this approach is if two partitions (of the two datasets) are exceedingly similar. A variety of measures have been proposed to remedy this problem. The remainder of this section provides an overview of the most common metrics.

4.2.3.1 Terminology

For precision, consider the following (formal) terminology and notation (*cf.* Meila (2007); Rosenberg and Hirschberg (2007)).

Let D be a set of N objects such that $D = \{d_a | a = 1, ..., N\}$.

A set of *clusters* $L = \{l_j | j = 1, ..., |L|\}$, where $|L|$ denotes the number of clusters, is a partitioning of the dataset D into disjoint subsets (called clusters), such that $l_j \cap l_m = \emptyset$.

A set of *classes* $C = \{c_j | j = 1, ..., |C|\}$, where $|C|$ signifies the number of classes, in turn, is a partitioning of the dataset D into disjoint subsets (called classes), such that $c_j \cap c_m = \emptyset$. Note that C is frequently also referred to as gold standard, as it represents the reference solution to a clustering task to which other clusterings are compared.

A clustering is *homogeneous* if every cluster only contains elements from a single class; it is *complete* if all elements of each class are assigned to the same cluster.

Clearly, the "perfect" outcome of a clustering exercise would be a (fully) homogeneous, complete clustering.

One commonly distinguishes three broad categories for the evaluation of clustering algorithms:

- **Mapping-Based Measures**.
  As the name suggests, in a post-processing step, these measures map each cluster to a class. Examples include L (Larsen and Aone, 1999), D (van Dongen, 2000), and the mis-classification index (Zeng et al., 2002). As these measures tend to be influenced by the mapping scheme used to map the clusters to the respective classes (Rosenberg and Hirschberg, 2007), they do not represent the ideal measures for evaluating my clustering methods and are henceforth disregarded.

- **Counting-Pair Measures**.
  These measures use a combinatorial approach that compares the number of pairs of data objects clustered similarly according to the gold standard and proposed clustering algorithm. Examples in this category include the Rand Index (Rand, 1971), Mirkin (Mirkin, 1996), and F-Measure (Hess and Kushmerick, 2003). Given their usefulness for the present work, the Rand Index and F-Measure will be described more formally below.

- **Information-Theoretic Measures**.
  As the techniques in this category evaluate full cluster membership (as opposed to mapped proportions only), they do not encounter the problems associated with mapping-based measures. Likewise, they evade the distributional problems associated with the counting pair measures. Information-theoretic measures are based on the notions of homogeneity and/or completeness, which in turn are expressed in probabilistic terms. Measures in this category include Mutual (Manning et al., 2008) and Normalized Mutual Information (Geiss, 2009), Variation (Meila, 2007) and Normalized Variation of Information (Reichart and Rappoport, 2009), the V (Rosenberg and Hirschberg, 2007) and V_{beta} Measures (Vlachos et al., 2009), Purity and Entropy (Zhao and Karypis, 2001), and Q (Dom, 2001). Given their usefulness for the present work, all but the last two measures will be described more formally below.

As a number of the foregoing measures are based on the correctness of the individual pair-wise memberships, consider the following notation:

TP $\equiv$ true positives. The two data objects belong to the same class and are assigned to the same cluster.

FP $\equiv$ false positives. The two data objects belong to different classes, but are assigned to the same cluster.

TN $\equiv$ true negatives. The two data objects belong to different classes and are assigned to different clusters.

FN $\equiv$ false negatives. The two data objects belong to the same class, but are assigned to different clusters.

(Conditional) entropies are denoted as follows:

$$H(C|L) = -\sum_{j=1}^{|L|}\sum_{i=1}^{|C|}\frac{n_j^i}{N}log\frac{n_j^i}{n_j}$$

$$H(C) = -\sum_{i=1}^{|C|}\frac{n^i}{N}log\frac{n^i}{N}$$

$$H(L) = -\sum_{j=1}^{|L|}\frac{n^j}{N}log\frac{n^j}{N}$$

4.2.3.2 F-Measure

The F-Measure, a counting-pair technique, is a widely used evaluation measure in information retrieval, first proposed by van Rijsbergen (1979). It is based on precision and recall, which it aims to combine such that the result is a measure for the accuracy of some algorithm relative to a gold standard. In the context of the evaluation of clustering algorithms, precision and recall are derived from pairs of objects as opposed to individual objects (as in Chapter 3), thereby circumventing the mapping from object to class, a problem that arises if the number of classes differs considerably from the number of clusters (Hess and Kushmerick, 2003). To be precise, precision is defined as $P = \frac{TP}{TP+FP}$ and recall as $R = \frac{TP}{TP+FN}$. The F-measure, in turn, is given by $F_1 = \frac{2 \cdot P \cdot R}{P+R}$. One of the main problems with the F-measure is that it is very sensitive to changes in the cluster partitioning, entailing that small changes in cluster assignments can have a substantial impact on the evaluation scores.

4.2.3.3 Rand Index (RI)

The Rand Index (Rand, 1971), also a counting-pair technique, is one of the earliest evaluation methods for clustering algorithms. It is defined as

$$RI = \frac{TP+TN}{TP+FP+TN+FN},$$

i.e., it derives the percentage of correct pair-wise decisions. The main weakness of this measure is that it does not utilize its (full) range between 0 and 1, but instead concentrates in a small interval near 1 (Meila, 2007). In addition, the pair-wise basis of this metric causes a single missing data object in a large cluster to have a much larger impact than a single missing data object in a small cluster.

4.2.3.4 Normalized Mutual Information (NMI)

The remaining measures in this section are more involved because of their roots in information theory. Mutual Information (MI) is based on the information that the gold standard and the clustering partitionings share and, using entropy and conditional entropy, can be expressed $MI = H(C) + H(L) - H(C,L)$.

A number of ways have been put forward to normalize the measure to ensure that the results are more easily interpretable. Manning et al. (2008), for instance, use the average of the two uncertainty coefficients (also *cf.* Press et al. (1988)), i.e.,

$$NMI = \frac{MI(L,C)}{\frac{H(L)+H(C)}{2}}.$$

4.2.3.5 Normalized Variation of Information (NVI)

When normalizing the Variation of Information (Meila, 2007), which measures completeness and homogeneity using conditional entropy, one obtains

$$NVI(L,C) = \frac{1}{logN}(H(C|L) + H(L|C)).$$

4.2.3.6 V and V_{beta}

The V-Measure (Rosenberg and Hirschberg, 2007) is based on homogeneity (h) and completeness (c) and is defined as

$$V(L,C) = \frac{(1+\beta)\cdot h \cdot c}{\beta \cdot h + c},$$

$$\text{where } h = 1 - \frac{H(C|L)}{H(C)} \text{ and } c = 1 - \frac{H(L|C)}{H(L)}.$$

Vlachos et al. (2009) proposes V_{beta}, where β denotes $\frac{|L|}{|C|}$, which automatically avoids the problem that the V-Measure favors clusters with much higher cardinality than the classes, a problem noted by Reichart and Rappoport (2009).

4.2.3.7 Discussion: Evaluation Metrics

This overview of metrics for the evaluation of clustering algorithms showed that a number of approaches are available and that significant effort has been expended on the improvement of the various methods. The idea of all of the approaches is that the evaluation metrics compare the clustering obtained by a given system to the (manually obtained) gold standard, and report the similarity between the two in the form of a score. In general, the best evaluation metrics are the information-theoretic approaches, since they avoid mapping problems and do not rely on pair-wise comparisons, which may result in inaccurate results because of the different impact of members of different clusters. For this reason, in my experiments, I evaluate clustering performance based on the NMI, NVI, and F-Measure metrics, the first two being selected on the basis of their theoretic foundation, while the last is chosen because of its wide-spread use in NLP evaluations.

4.3 Fully Automated Derivation of a Pyramid

The fundamental problem with the automatic acquisition of a Pyramid-style evaluation framework – as has been suggested when constructing the partially automated variant (*cf.* Chapter 3) – is the considerable variability of the realization of similar or near-identical ideas in the reference summaries. Other automatic summarization evaluation methods deal with this problem by accounting for the word overlap between two summaries, e.g., ROUGE (Lin and Hovy, 2003). The main shortcoming of this approach is the reliance on the surface realizations of the informational content. In addition, ROUGE only accounts for relations between words in a very limited manner, namely, by considering n-grams. In order to amend these limitations, as before, I distinguish between the underlying "concept" (or idea) of the informational content, the syntactic realization of the sentences, and the entity/event realizations in the reference summaries. As a result, a single concept can have multiple syntactic realizations,

as can the entities in the concept. For reference, my representation of an individual underlying concept is reproduced in Figure 4.2.

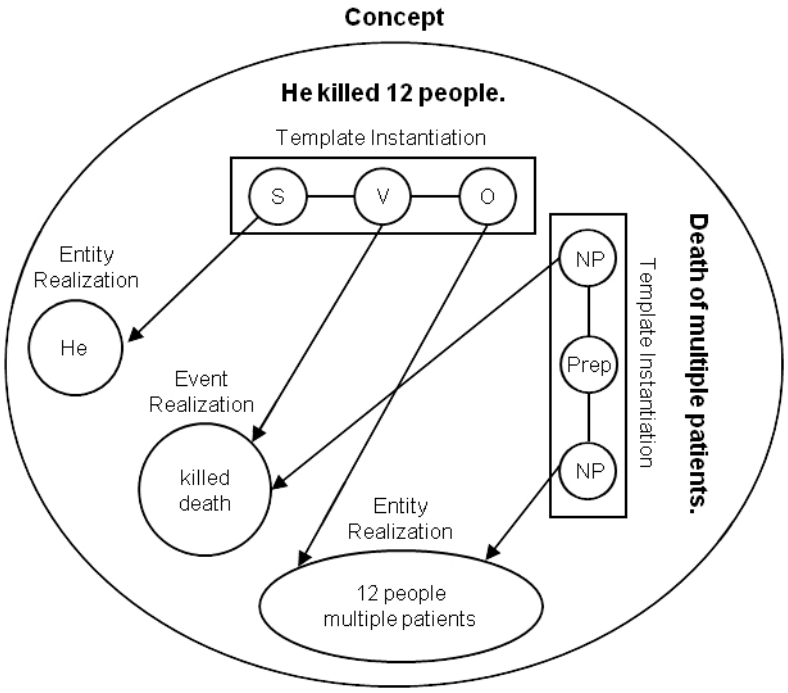

Figure 4.2: The "Concept." The relationship between underlying concept, syntactic realization, and entity/event realization; reproduced from Chapter 3.

The concept representation forms the basis for the creation of the fully automated, Pyramid-style evaluation framework. I use an agglomerative, hierarchical clustering algorithm to continually combine concepts until there are no further clusters that are sufficiently similar to one another. To illustrate the process, Figure 4.3 provides pseudo-code for the clustering algorithm. In a first step, for each sentence fragment corresponding to a template (i.e., a template instantiation), an individual concept is created. Then, a pair-wise similarity matrix is generated, the details of which are discussed in Section 4.4. In a third step, the two most similar concepts are combined, until no two concepts are sufficiently similar based on a manually determined threshold (obtained using partial exploration of the threshold space). The result is an automatically generated pyramid from the (manual) reference summaries. The remainder of this section details this outline of the proposed procedure.

```
//Step 1
concepts = createConcepts(document collection)
//Step 2
similarityMatrix[][]=
    createPairwiseSimilarityMatrix(concepts)
int column, row; double max = 0;
do {
 iterate over all field[i][j] in the matrix {
  if (field > max)
   max = field; column = j; row = i;
 }
 if (max >= THRESHOLD) {
  combineConcepts(i, j, concepts);
  similarityMatrix =
    updateMatrix(similarityMatrix, i, j, concepts)
 }
} while (max >= THRESHOLD)
//Step 3
cooccurrenceMatrix[][] =
    constructPairwiseCooccurrenceMatrix(concepts)
int column, row; double max = 0;
do {
 iterate over all field[i][j] in the matrix {
  if (field > max)
   max = field; column = j; row = i;
 }
 if (max >= THRESHOLD2) {
  joinConcepts(i, j, concepts);
  cooccurrenceMatrix =
        updateMatrix(cooccurrenceMatrix, i, j, concepts)
 }
} while (max >= THRESHOLD2)
```

Figure 4.3: Hierarchical Clustering of Concepts.

4.3.1 Algorithmic Approach

The idea underlying all of the research presented in this thesis is that a limited set of syntactic relationships along with lexical semantic information is useful to capture the salient similarities in informational content. Rooted in this notion, in the context at hand, the (human) reference summaries need to be converted to the concept representation using the templates introduced in the previous chapter. This process is captured by the first line of pseudo-code ("Step 1") in Figure 4.3.

The second stage in the automatic creation of the pyramid is the grouping of similar template instantiations. As clustering is the task of assigning a set of data objects into subsets such that the data objects in the subsets are similar according to some metric, it is clearly appropriate for this stage. With regard to the particular clustering approach to be used, a number of aspects need to be taken into account. First and foremost, it was found in Chapter 3 that sharing information between template instantiations (by combining them into concepts) was beneficial for the overall matching process. For this reason, it would be preferable if the clustering approach would exploit all of the information in a given cluster as opposed to metrics relying on similarity of the individual data objects only. In consequence, as stated, density, partitioning, and grid-based clustering approaches are problematic, as they require an assumption relating to the independence of the data objects. The presented algorithm for creating the pyramid therefore uses a hierarchical clustering approach. The only requirement for hierarchical clustering is the feasibility of computing similarity measures between two data objects, or, in a more general case, between clusters.

Even though, in principle, single, average, or maximum similarity linkage metrics are applicable, they do not exploit the information that can be gained by sharing information between templates. Conceptual clustering, on the other hand, allows for the exploitation of this information by explicitly representing the information from all data objects in the cluster description. Note that this linkage metric in conjunction with an (exclusive) agglomerative, hierarchical clustering approach allows for updating the cluster description at each step. As a result, this combination of features is precisely what is used for the second step: the linkage metric uses the concept representation developed in Chapter 3 to represent clusters. Every time two clusters in the clustering process are combined, so are the two concept representations, and the relevant rows and columns in the similarity matrix are recomputed. These steps are represented by "Step 2" in Figure 4.3.

"Step 3" relates to the use of variable-sized informational units. Although, in "Step 2," the templates are clustered together, it does not make use of the co-occurrences of template instantiations, i.e., the fact that two template instantiations occur in the same sentence most of the time. That is to say, one of the most essential advantages of the manual Pyramid evaluation method over other approaches is not exploited. In the Pyramid scheme, the size of SCUs is determined via the co-occurrence of information in the reference summaries. If all summaries contain two particular pieces of information, then these two pieces form a *single* SCU. For a concrete example, refer to Figure 2.13 in Chapter 2, and consider the occurrence of "1993" in C6, E1, and F3. If "1993" also occurred in the other sentences containing SCU1, then the fact that it was in 1993 that Lopez left GM for VW would be included in SCU1. In order to use the same principle in the automated method for constructing a pyramid, in a third stage, clusters are joined into SCUs based on their co-occurrence statistics. The result is an automatically generated pyramid of SCUs based only on a (limited) set of manual reference summaries.

The threshold parameters in the second and third step once more derive from a series of runs of the system using different parameter values, in steps of 0.05, and the performance of the system using these different settings. In this manner, the optimal value of "THRESHOLD" parameter (template similarity) was determined to be 0.55 and that for the "THRESHOLD2" parameter (cluster overlap) to be 0.65.

Note that the second and third steps are ultimately both clustering stages. However, they cluster the concepts based on very different criteria. While the second stage (hereafter referred to as "clustering concepts") clusters the concepts based on syntactic and semantic similarity, the third stage clusters the concepts based on the co-occurrence statistics between the individual concept clusters (hereafter referred to as "cluster composition" or "joining clusters"). In other words, the information is first clustered such that similar information is combined, whereupon information that is frequently close to each other is grouped together. The result is a set of variable-sized units of information.

4.3.2 The Task and Manual Document Annotation

The algorithmic approach just presented is based on the instructions for manual Pyramid annotation; Appendix D provides the full annotation instructions. While the task to be achieved by the semi-automatic system introduced in the preceding chapter only required identifying whether (or not) an SCU is present in a specific sentence, the fully

automatic version proposed here also requires the identification of the size of a unit of information in addition to the identification of similarity. To achieve this aim, a look at the instructions for the manual annotation provide helpful insight.

The annotation instructions state that an SCU is generally no larger than a clause. The proposed approach, in principle, incorporates this constraint via the templates exclusion list (*cf.* Chapter 3.3). However, owing to the syntactic structure of complex expressions, my approach does not comprise a hard requirement that limits the SCU to a single clause. In practice, there are very few situations in which problems occur due to this design choice. It is stated that the size of the SCU is typically determined by the information overlap between the different summaries in which the SCU occurs. This implies that one first identifies the main information of the SCU in the different summaries and subsequently determines adjuncts that belong to the SCU.

One can make out several similarities between the manual annotation and the automated approach presented above. First, common main units of information in the different summaries are identified (Step 2), and then adjuncts that belong to this information and occur in all reference summaries are determined (Step 3). Thus, the approach presented in this section follows the general instructions given to the human annotators.

The experiments presented in the next section develop the actual implementation for the high-level processes underlying the various processing steps presented in this section. The first set of experiments explores the initial clustering based on template information only. The second experiment investigates the usefulness of contextual information for the clustering process. Subsequently, the use of the concept representation and the joining of clusters are explored. For the penultimate experiment, I combine the automation of the second step of the (original) Pyramid Evaluation scheme (*cf.* Chapter 3) and the automatically generated pyramid developed in this chapter in order to obtain a fully automatic evaluation method for the informational content of peer summaries, and compare the rankings obtained using my approach, ROUGE, and Harnly et al. (2005)'s partial pyramid automation to the rankings obtained using the manual Pyramid method. The final experiment automatically evaluates automatically generated summaries in the TAC 2009 AESOP dataset, which provides a useful comparison to other recently developed evaluation systems.

4.4 Experiments

From a conceptual point of view, the key question remaining regards the details of the similarity function to create the appropriate connectivity matrix, which – from a theoretical perspective – needs to incorporate and/or account for the following considerations:

- the similarity of all relevant syntactic constituents in the clusters (closeness in terms of WordNet);
- the similarity of the context in which the clusters occur;
- the appropriateness of the syntactic transformations required for the clusters;
- the combination of the information from different surface realizations for the same cluster (information sharing as in the SCU matching stage);
- the fact that only one unit in a document can belong to one particular cluster (based on the simplifying assumption that a document does not contain duplicate information); and
- the number of other members already contained in a given cluster.

While all of these attributes are intuitive, this section investigates the practical use of each of them. To this end, for transparency, the problem is divided into several stages: Initially only template instantiations are clustered, then contextual information is exploited in the clustering process, then information sharing, and finally cluster composition.

All but the last two experiments are evaluated on the 10-cluster development dataset used for the initial experiments in Chapter 3. The penultimate experiment, comparing performance of the fully automated evaluation process, is evaluated on the remaining 40 clusters in the dataset. To gauge the system's performance in a broader context, the final experiment is based on the TAC2009 AESOP dataset.

4.4.1 Experiment 1: Initial Clustering Using Template Information

The first experiment only considers the similarity of the syntactic constituents, and appropriateness of the syntactic transformations. In its substance, the experiment is a modified version of the first experiment reported in Chapter 3. In particular, it contrasts

a number of different methods to achieve word-based matches, both in the context of unstructured and syntactically structured clustering, i.e., using the syntactic information in the clustering process (or not). The details of the various matching methods are identical to those described there. The clustering function for these experiments is simplistic. In view of the experiments' objective, it determines the similarity between two template instantiations (or the words in the template instantiations in the case of "None") based on the similarity of the relevant constituents of the two template instantiations. The results are summarized in Table 4.1.

Syntactic Structuring	**Word-based Matching**	**NMI**	**NVI**	**F-Measure**
None	lemma + c. + s.	0.40	0.45	0.50
None	WordNet + c. + s.	0.45	0.47	0.47
Basic Trio	lemma + c. + s.	0.55	0.57	0.53
Basic Trio	WordNet + c. + s.	0.59	0.59	0.57
Trio + NPPrepNP + XPrepX	lemma + c. + s.	0.56	0.57	0.55
Trio + NPPrepNP + XPrepX	WordNet + c. + s.	**0.60**	**0.61**	**0.60**

Table 4.1: Results of Experiment 1. Evaluation of clustering sub-sentential units based on syntactic structure and word-based similarity measures. "Basic Trio" denotes the combination of the SVO, SV, and VO templates. "c." stands for content words and "s." for stop word list.

They demonstrate that the use of more varied syntactic templates improves the performance of the clustering algorithm as long as only the main syntactic transformations are considered (i.e., SVO + SV + VO + NPPrepNP + XPrepX). In particular, further results (not shown) did not show improvements if modifier relations are considered. Similarly, the addition of WordNet improves the results in all syntactic situations, except when no syntactic templates are used. Intuitively, this difference is easily explained on account of the over-generalization of WordNet relations in conjunction with the missing constraints imposed by the syntactic templates.

4.4.2 Experiment 2: Clustering Using Contextual Information

In the next experiment, I consider the impact of the context of the syntactic templates. For example, based on the SV template, one might obtain two instantiations containing "he killed," but in one situation the killing in question occurs in, say, 1994, while the

other does not occur until 2006. Clearly, even though the instantiations are identical, they should not be clustered together since they express different content. To moderate such situations, I incorporate the context of the instantiations into the clustering function. In particular, the experiment investigates four different approaches: (1) word-based overlap similarity between the two contexts; (2) overlap between named entities in the two contexts; (3) similarity of the syntactic instantiations in the two contexts; and (4) the combination of (2) and (3). The first approach clearly establishes a baseline for the utility of considering the context of template instantiations, while the second and third explore different aspects of contextual information on the clustering process.

A manual inspection of the SCUs created for the summarization evaluation revealed that there are SCUs that mainly act as "helper" SCUs because they are clearly related to the main SCU, but not all summaries contain the additional information. Many of these helper SCUs contain named entities and dates or places. This discovery formed the basis for the second approach, since analogous named entities in the context of a template instantiation increase the likelihood of the similarity of the syntactic instantiations, while non-contradictory absence of such similarity does not necessarily constitute a mismatch. Rather, contradictory named entities definitely decrease the likelihood of a match. Coming back to the dates in the killing examples, the mention of 1994 in one and 2006 in the other without doubt means that they do not cover the same event. The presence of a date in one while the other lacks a date, however, only means that one does not have sufficient information to resolve the situation. The presence of the same date in both situations, in turn, indicates a strong likelihood that the two instantiations concern the same event. The third approach is based on the same idea as the syntactic templates, i.e., the importance of the relationships between words and/or entities. As such, in this approach, I consider the percentage overlap of template instantiations in the contexts of the template instantiations under consideration. A high overlap implies a higher probability that two instantiations are to be clustered together. The fourth approach investigates whether a combination of the two preceding approaches improves recognition, or whether one of the approaches by itself is preferable.

The results of the investigation of these four hypotheses are presented in Table 4.2. They show that accounting for context similarity clearly improves performance, though the most significant gain is obtained when using both named-entity and syntax-based similarity measures. The improvement when combining the measures, however, is minimal.

Similarity Measure	NMI	NVI	F-Measure
None	0.60	0.61	0.60
Word-based Similarity	0.63	0.65	0.63
NE-based Similarity	0.65	0.66	0.64
Syntax-based Similarity	0.64	0.66	0.65
NE + Syntax-based Similarity	**0.67**	**0.68**	**0.66**

Table 4.2: Results of Experiment 2. The impact of context similarity on clustering performance. "NE" stands for named entity.

4.4.3 Experiment 3: Clustering Using Concepts

Having investigated the impact of contextual information on the validity of the matches between different syntactic instantiations, I now turn to the issue of integrating the information from multiple identified matches. That is, if two constituents were identified to refer to the same underlying entity or relation, an additional performance gain can be obtained by exploiting this information as input for future processing steps (*cf.* Section 3.6.4 for more details). On the road to achieving this end, the proposed approach combines all syntactic instantiations identified as similar into a single data structure called concept, i.e., it shares information between different template instantiations (for a more detailed description of information sharing, cf. Chapter 3).

The results of using these partial concepts are displayed in Table 4.3. While they reveal a slight gain in overall performance, the results are (still) rather poor. At this point, note that the clustering algorithm only grouped individual syntactic instantiations. That is, referring back to Figure 4.1, for the purposes of evaluation, the three syntactic instantiations in the SCU with ID 14 are all in the same class. Therefore, the syntactic and semantic similarity that has thus far been exploited is not capable of determining these problems. The following experiment addresses this issue.

Method	NMI	NVI	F-Measure
No Information Sharing	0.67	0.68	0.66
Information Sharing	0.69	0.71	0.67

Table 4.3: Results of Experiment 3. The impact of using partial concepts in the clustering of syntactic instantiations.

4.4.4 Experiment 4: Cluster Composition

The penultimate experiment corrects the foregoing shortcoming by combining clusters of single syntactic instantiations into clusters containing multiple different syntactic instantiations in a post-processing step. To this end, two clusters are combined (only) if, in 75% of the cases, the two clusters are both contained within the same sentences. The results of this experiment are presented in Table 4.4. They exhibit a significant (though expected) increase in the clustering evaluation measures. In addition, the results now indicate that the automatically generated clusters are quite similar to the clustering of template instantiations based on the SCU annotations from the manual evaluation because of the normalized nature of the evaluation metrics; that is, the scores approach the upper bound which indicates that the clusters are very similar. On the whole, this experiment shows that the use of the two-stage clustering approach – using information sharing *and* contextual information – presented in this chapter results in similar clusters as the manual Pyramid evaluation scheme.

Method	**NMI**	**NVI**	**F-Measure**
No Cluster Composition	0.69	0.71	0.67
Cluster Composition	0.89	0.90	0.89

Table 4.4: Results of Experiment 4. The impact of combining clusters of individual syntactic instantiations based on proximity constraints.

4.4.5 Experiment 5: Evaluation of the Fully Automated Pyramid-Style Method

So far, I only evaluated the similarities of the clusterings obtained by using the presented algorithms to the gold standard clustering derived from the Pyramid information. However, a very similar clustering evaluation measure does not guarantee that the (final) scores achieve good performance in the overall automatic summarization evaluation task, as measured by comparing the rankings produced by the evaluation methods. To measure the effectiveness of my automated version of the Pyramid evaluation scheme (i.e., the combination of the algorithms presented in this and the preceding chapter), I present the results of the ranking correlation (tau) of the manual Pyramid method against the ROUGE evaluation measure, Harnly et al. (2005)'s word-based

clustering, and my measure. In particular, I first obtain a pyramid using the clustering algorithms presented in this chapter and then match the pyramid into the system summaries using the approach presented in Chapter 3.

The results of this penultimate set of experiments are displayed in Table 4.5. They leave no doubt that the present procedure outperforms both ROUGE and Harnly et al. (2005). The results also indicate that despite Harnly et al. (2005)'s and Lin (2004)'s evidence that their methods cannot be improved when using syntactic relations, my approach does achieve this goal.

Method	**Ranking Correlation**
Partial Automation of Pyramid Approach (Chapter 3)	0.97
Full Automation of Pyramid Approach	0.96
ROUGE-2	0.93
Word-Based Clustering (Harnly et al., 2005)	0.95

Table 4.5: Results of Experiment 5. (Spearman ρ) Ranking correlation of different methods against the original manual Pyramid evaluation method.

I would claim that this improvement in performance is likely to be caused by two design choices. First, I limit the type of syntactic relations considered to be important to a small number of syntactic relations, for which intuitive ideas of their use and usefulness are provided. Second, by using knowledge sources – WordNet in particular – I provide for a greater number of syntactic matches because of the use of relations between syntactic constituents beyond the surface form or lemma similarity.

4.4.6 Experiment 6: Evaluation Using AESOP2009 Dataset

Paralleling Chapter 3, the final experiment in this chapter compares the performance of the *fully* automated Pyramid evaluation to twenty-four competing system using the TAC2009 AESOP dataset (*cf.* Section 3.6.6). In contrast to that experiment, here I only use the information officially available for the evaluation systems, i.e., the manually created pyramids are no longer used.

Tables 4.6 and 4.7 compare the present system to other systems in terms of the correlation of their rankings with the manual Pyramid method, when using the summaries for the initial documents and the update summaries, respectively. Only NoModels results are reported because the NoModels and AllPeers evaluations are very similar and

System ID	**Pearson** (r)	**Spearman** (ρ)	**Kendall** (τ)
6	0.911	0.950	**0.820**
12	0.901	0.947	0.815
26	**0.978**	0.942	0.810
15	0.805	0.939	0.800
Full Automation	0.920	**0.952**	0.799
11	0.954	0.933	0.796
10	0.869	0.931	0.793
2 ROUGE-BE (baseline)	0.857	0.936	0.791
4	0.967	0.928	0.788
25	0.850	0.928	0.787
1 ROUGE-SU4 (baseline)	0.921	0.923	0.785
13	0.952	0.924	0.785
28	0.830	0.933	0.785
18	0.965	0.918	0.770
19	0.967	0.917	0.769
23	0.928	0.908	0.765
17	0.952	0.908	0.759
31	0.894	0.912	0.758
5	0.799	0.915	0.757
32	0.815	0.901	0.755
33	0.742	0.900	0.752
24	0.963	0.902	0.750
16	0.962	0.900	0.742
21	0.796	0.887	0.730
34	0.897	0.873	0.715

Table 4.6: Results of Experiment 5 (Part 1). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for NoModels when using the initial summaries. The results are sorted according to Kendall's τ metric. Systems with ID 1 and 2 are the ROUGE-SU4 and ROUGE-BE baselines.

System ID	**Pearson** (r)	**Spearman** (ρ)	**Kendall** (τ)
28	0.908	**0.955**	**0.841**
15	0.887	0.950	0.831
Full Automation	0.931	0.948	0.822
2 ROUGE-BE (baseline)	0.924	0.932	0.801
25	0.896	0.937	0.800
16	0.968	0.918	0.789
10	0.918	0.924	0.785
12	0.946	0.923	0.781
24	0.957	0.916	0.781
19	0.962	0.911	0.781
23	0.932	0.912	0.776
5	0.895	0.920	0.774
11	**0.970**	0.918	0.772
18	0.944	0.901	0.768
26	**0.970**	0.903	0.768
33	0.855	0.919	0.768
13	0.962	0.904	0.754
6	0.921	0.902	0.753
8	0.937	0.880	0.734
31	0.940	0.884	0.734
4	0.946	0.874	0.719
1 ROUGE-SU4 (baseline)	0.940	0.863	0.708
21	0.652	0.848	0.702
17	0.944	0.847	0.698
27	0.934	0.854	0.695
34	0.767	0.853	0.683

Table 4.7: Results of Experiment 5 (Part 2). Ranking correlation with manual Pyramid evaluation scheme on the TAC2009 AESOP dataset for NoModels when using the update summaries. The results are sorted according to Kendall's tau.

NoModels provides the more challenging task. The results show that my fully automated approach performs rather well, ranking 5^{th} for the initial summaries and 3^{rd} for the update summaries. More specifically, my approach is only outperformed on both datasets by System 15, while the other systems only rank higher for one of the systems.

4.5 Remarks Relating to Statistical Significance and Confidence Intervals

Statistical significance and/or confidence intervals can generally be used to determine whether the evaluation results obtained by a given system differ from those achieved by other systems or human judges. In the context of the (official) results from the AESOP-2009 evaluation, however, a visual inspection reveals that the p-values (signifying statistical significance between the human judgments and the system judgments) do not provide any useful insight into their relative performance apart from cases involving the *worst*-performing systems.

When considering confidence intervals, the problem is similarly situated. The intervals are usually too wide to make any but the coarsest distinctions; the confidence intervals associated with the different systems tend to overlap heavily. The common width of the relevant intervals is around 0.1, with the narrowest interval being 0.025, for Pearson's correlation coefficient on the initial summaries.

4.6 Sample Passages Highlighting Strengths and Weaknesses of my Automatic Evaluation System

Having established that the proposed fully automatic Pyramid-style evaluation system compares well to the majority of competing systems, the purpose of this section is to highlight some of its merits and weaknesses by way of two sample text passages taken from DUC2005 documents. As before, they are selected to illustrate situations in which the approach identifies information well, as well as scenarios in which it fails to find any similarities. In fact, they are the same as those Section 3.7 in the preceding chapter. The first sample passage, presented in Figure 4.4, constitutes an example for which the proposed clustering mechanism works well.

In general, the syntactic structure of the different contributors is very similar: Subject {Malaysia, it, South East Asia's Malaysia} – Verb {used to be, was, became, was

```
<scu uid="252" label="Malaysia used to be the world's premier producer of tin">
 <contributor label="Until recently, it was the world's premier producer">
  <part label="Until recently, it was the world's premier producer" start="901"
   end="952"/>
 </contributor>
 <contributor label="Until recent times, Malaysia was the world's largest producer
  of tin">
  <part label="Until recent times, Malaysia was the world's largest producer of tin"
   start="1641" end="1709"/>
 </contributor>
 <contributor label="Malaysia became the world's leading tin producer">
  <part label="Malaysia became the world's leading tin producer" start="3601"
   end="3649"/>
 </contributor>
 <contributor label="Malaysia was once the world's leading tin producer">
  <part label="Malaysia was once the world's leading tin producer" start="4704"
   end="4754"/>
 </contributor>
 <contributor label="Malaysia gradually became the leading producer of tin in the
  world">
  <part label="Malaysia gradually became the leading producer of tin in the world"
   start="6350" end="6416"/>
 </contributor>
 <contributor label="Since 1857 South East Asia's Malaysia had been the world's
  largest tin producer">
  <part label="Since 1857 South East Asia's Malaysia had been the world's largest
   tin producer" start="7922" end="8001"/>
 </contributor>
</scu>
```

Figure 4.4: Example 1. Representative portion of the manual Pyramid analysis of document cluster D632 (DUC2005)

once, gradually became, had been} – Object {the world's premier producer of tin, the world's largest producer of tin, the world's leading tin producer, the leading producer of tin in the world, the world's largest tin producer}. The only difficult part in the Subject constituent is the pronoun, for which a match to Malaysia is determined based on the possible antecedents given the preceding sentence and the syntactic similarities between the members of the SCU. A more interesting problem is encountered when looking to match the verbs. WordNet does not identify a relation between "be" and "become" and it is therefore impossible to relate the two verb groups directly. However, since template similarity does not have to be exact, the match can be found by way of the equivalence of the Subject and Object in the third and fourth contributors. Using information sharing, at a later stage, phrases such as "It became the largest producer of tin in the world" can ultimately be identified very accurately owing to the overlap between the different entities/events in the concept data structure.

Besides matching the main information in the different sentences ("Malaysia was a producer"), the clustering procedure also works well with respect to matching the relevant adjuncts ("producer of tin" and "world's leading producer"). Note that each of the contributors contains either "tin producer" or "producer of tin." Thus, attaching this adjunct to the main unit of information is, in fact, quite straightforward because the co-occurence of tin producer with the Subject – Verb – Object relation is high. The situation is similar for "the world's leading producer," though in this case, the variation is slightly higher ("world's premier producer," "world's largest producer," "leading producer in the world"). As they can be reduced to the syntactic variations in the NPrepN category, they too can successfully be attached to the main unit of information. Correspondingly, apart from some adjunct information that does not actually occur in all contributors ("until recent times" and "since 1857"), the algorithm successfully identifies the complete SCU. Note, in this context, that according to my understanding of the annotation instructions, these adjuncts should technically constitute individual SCUs since they introduce significant information, i.e., that Malaysia is no longer the largest producer, and when it first became the largest producer. As this is not the case, their omission by the matching procedure is not necessarily detrimental.

The second representative portion of text, presented in Figure 4.5, is an example of an SCU for which the approach does not yet perform well. In particular, the size and form of the two last contributors cannot be obtained by the approach in this chapter. Looking at the full sentences for the relevant instances, one can see that an appropriate unit of information that could potentially be identified is "to be trained to identify [..]

narcotics." Moreover, using information sharing, "sniff" and "detect" can be identified to be similar. Yet, the association of training and the detection of narcotics does not occur in the algorithm because the co-occurrence between the two facts is too small. As such, the algorithm nonetheless identifies an SCU with 5 contributors instead of 6. The last one is not identified because: (1) there is no link between "search," "detect," and "sniff," and (2) the syntactic structure for the sentence is not correctly identified, so that it is impossible to induce similarity based on the syntactic structure.

```
<scu uid="146" label="Dogs are used to sniff out narcotics">
 <contributor label="to sniff narcotics">
  <part label="to sniff narcotics" start="443" end="461"/>
 </contributor>
 <contributor label="Dogs trained to sniff out narcotics">
  <part label="Dogs trained to sniff out narcotics" start="2388" end="2423"/>
 </contributor>
 <contributor label="their use in detecting narcotics">
  <part label="their use in detecting narcotics" start="4955" end="4987"/>
 </contributor>
 <contributor label="to sniff out evidence of drugs">
  <part label="to sniff out evidence of drugs" start="6629" end="6659"/>
 </contributor>
 <contributor label="and have been trained to detect narcotics">
  <part label="and have been trained to detect narcotics" start="10461" end="10502"/>
 </contributor>
 <contributor label="drugs">
  <part label="drugs" start="8576" end="8581"/> (full sentence: The tasks performed
   by dogs include searching passengers, vehicles, aircraft, ships and cargo for
   bombs, drugs and agricultural contraband at borders, ports, and airports, as well
   as at crime scenes or on regular patrol.)
 </contributor>
 <contributor label="narcotics">
  <part label="narcotics" start="3562" end="3571"/> (full sentence: A dog's keen
   sense of smell enables it to be trained to identify many types of narcotics,
   explosives, and flammable material.)
 </contributor>
</scu>
```

Figure 4.5: Example 2. Representative portion of the manual Pyramid analysis of document cluster D426 (DUC2005)

4.7 Discussion

The work presented in this chapter investigated the steps necessary to develop a fully automated version of the Pyramid evaluation method, i.e., the derivation of a pyramid

of semantic content using a hierarchical clustering approach. The experiments showed that both WordNet semantic knowledge and specific types of syntactic relations provide a good basis for such an endeavor. By using contextual information, information sharing between syntactic instantiations in so-called concepts, and by combining these concepts where possible via proximity constraints, I was able to obtain clusters of information that are highly similar to those derived when using the information available via the original manual Pyramid. Last but not least, I showed that the components for the partial and full automation of the Pyramid scheme work well in concert. Namely, they result in rankings that are more similar to the manual Pyramid evaluation method than a non-negligible number of other state-of-the-art automatic evaluation methods.

Chapter 5

Summary Generation Using Variable-Sized Informational Units

5.1 Introduction

The purpose of this chapter is the construction of a methodology to *generate* summaries from multiple document sources. One of the central aspects in this regard is the selection of the appropriate informational content from the underlying sources. To arrive at a workable solution for this task, the procedure proposed in the following exploits several elements of the techniques developed as part of the automatic evaluation method for system generated summaries presented in the preceding chapters.

For the reasons outlined at the outset of this thesis, summarization has, in recent years, received a lot of attention within the NLP community. The solutions put forward have been varied, ranging from very simple designs based on word frequencies to full-scale logic-based approaches. The main advantage of systems based on word frequencies is that one can easily apply mathematical and statistical models in order to optimize sentence selection. Given their simplicity, however, they do not account for relations between individual words, which is a notable disadvantage because natural language contains more than just words. The relationships between words are of vital importance.

The objective of this chapter is to extend the word-frequency methods to encompass word relations and determine individual content units, that is elemental units of information. The main characteristics of the proposed approach are: (1) the individual units of information to be included in the summary vary in size; (2) units that occur more often in the source documents are considered more important; and (3) the summarization approach is based on the creation of a pyramid of content units (*cf.* Chapters 3 and 4). In order to be able to apply the relevant techniques underlying my Pyramid-style automatic evaluation method successfully to the automatic generation of summaries, it is necessary to consider the similarities and differences between the two scenarios. To this end, the following list summarizes the main resemblances and disparities with respect to the foundations of the relevant systems:

Similarities

- Both processes are based on informational units.
- The informational units are variable in size.
- Similar informational units are clustered together.
- The frequency of informational units across the source documents is taken to be

an indicator of importance.

Differences

- The "type" of underlying document is fundamentally different. In cases of summarization, the inputs to the systems are the documents to be summarized, while evaluation methods are based on human reference and system summaries.[1]
- The length of the documents to be summarized tends to be considerably greater.
- Frequency of information is not the only indicator of importance for summarization.

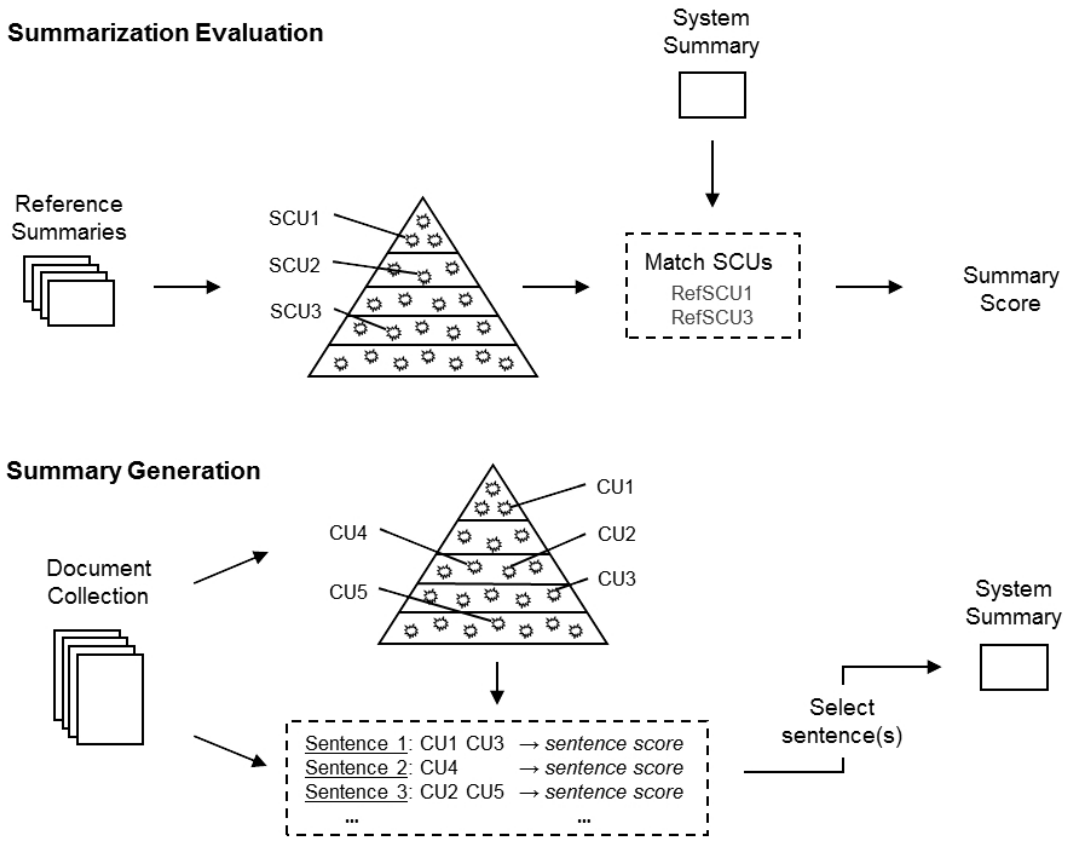

Figure 5.1: The differences between the summarization and evaluation processes using variable-sized informational units.

Even though the summarization and evaluation systems of interest in the present context are based on the same underlying idea – the use of variable-sized informational units – Figure 5.1 graphically summarizes the key differences in the underlying

[1] In practical terms, this implies that the input to evaluation methods is slightly more topic-specific than the documents to be summarized, which are usually only loosely associated with one another.

processes. In particular, it illustrates the dissimilar use of the content units (that are created using a Pyramid-style approach) in the summary generation task when compared to their use in the evaluation scenario. In the summary generation process, the content units are used to obtain scores for each of the sentences in the source documents based on the properties and frequency of the content units. These sentence scores are then used to select the sentences to be included in the summary text; note that the sentence scores change based on the sentences already selected for the summary. As described in detail in previous chapters, in the evaluation scenario, the content units identified in the human reference summaries are matched to those also occurring in the system summary. The degree of "overlap" is then used to compute a score for the quality of the summary's informational content relative to the content of the reference summaries. Hence, in a nutshell, the fundamental difference between the overall processes is the different use of the content units.

In view of these considerations, the work presented in this chapter makes two main contributions to the field of automatic multi-document text summarization:

- it proposes an approach that allows for the extraction of variable-sized units of informational content from the source documents (this chapter focuses on newswire documents, but the approach should apply to wide variety of domains); and

- it demonstrates that the Pyramid annotation scheme is a valuable means for the development of summarization systems, which opens up a new field of use for evaluation annotations.

The remainder of this chapter is organized as follows. Section 5.2 presents and contrasts the work motivating and informing the proposed methodology. Section 5.3 subsequently describes my summarization approach in more detail and illustrates the workings of the algorithm by way of an example. Sections 5.4 and 5.5 describe my method for determining which surface representation to include in the summary text and how I deal with the potential issue of repetitive information, respectively. Section 5.6 outlines the experimental results when testing the system on a number of DUC datasets. Section 5.8 illustrate the workings of the sentence selection algorithm based on a number of examples, that also highlight some of the problems encountered during the sentence selection process. Section 5.9 concludes the chapter with a brief discussion of the main results.

5.2 Related Work

Systems based on word-frequency measures are among the most common approaches to automatic summarization. The main difference between the various systems in this class is how the importance of individual words is transferred to sentences or sub-sentence units to be included in the summary text. One of the earliest systems in this context is that by Luhn (1958), who "used statistical information derived from word frequency and distribution [...] to compute a relative measure of significance."

A more recent, quite successful, example is SumBasic (Nenkova et al., 2006), who utilize the probabilities of the words in the original documents (to be summarized) in order to obtain the most likely summary; the authors apply the basic rules of probability theory to the words and their occurrence in the summary. The result is an intuitive way to deal with duplicate words within selected sentences, as the probability of a word occurring twice is simply the square of the original probability.[2] While this approach is very interesting and avoids the duplication of summary content without additional post-processing steps, the underlying idea does not translate well into the present summarization framework. For, the informational units of interest are significantly larger and, more importantly, as the proposed approach explicitly considers the similarity of textual (content) units, it does not require a separate method for dealing with duplicate information as an informational unit should only be selected for a summary at most once.

Leskovec et al. (2004) extend a word-frequency approach so as not to focus on individual words. Instead, they employ word-triples to capture shallow relations between words, which they merge into a semantic graph on the assumption that every word in a given collection of documents has a single sense/meaning. On the basis of this semantic graph, they compute a number of features and learn a (binary) classifier to extract sub-trees, which ultimately represent the summaries. The main feature of their system is a PageRank score, according to which each unit in a triplet corresponds to a webpage in the original PageRank task and each link between the units of the triplet corresponds to a link between two webpages. In short, Leskovec et al. (2004) combine a semantic graph derived from word-triples with several linguistic features to select a sub-graph to represent a summary.

[2]The probability of two idependent events occurring at the same time is given by the product of the individual probabilities. Hence, in the case of the *same* event occurring twice, the joint probability is simply the square of the probability of the event occurring once. (Nenkova et al., 2006) assume independence between the occurrences of the different words.

Note, however, as do Leskovec et al. (2004), that the use of triplets to select sentences is not necessarily the best option for creating a summary, as usually more information than required by the triplets is selected for the summaries. To alleviate this problem and achieve more accurate summaries, one might use natural language generation methods on specific semantic graph segments to create abstractive summaries. As their (original) system already identifies the most relevant sections of the source documents, generating full sentences based on the selected sub-graph might be a better, though a significantly more complex, approach.

The approach by Leskovec et al. (2004), amongst other things, illustrates the usefulness of syntactic information, i.e., the use of information regarding the relationships between words, for the summarization approach. In other words, it is to be expected that by combining frequency-based approaches with syntactic information, one can develop summarization systems that extract relevant information more accurately. On the note of triplets, the selection of sentences as described in the previous paragraphs and as is common in summarization research is by no means the only option. Even ignoring natural language generation potential for the realm of abstractive summarization, which is concerned with paraphrasing text from one or more document sources, the following options are available in the realm of extractive summarization:

- **Sentence Extraction.**

  The extraction of full sentences from the source documents (to be included, as is, in the summary text) is the most common method used in automatic summarization research. A score or a collection of scores describing their importance is computed for each sentence in the original documents. These scores form the basis for determining which sentences should be included in the summary text.

- **Discourse-Unit Extraction.**

  The starting point for this approach is the observation that many sentences are constructed using coordination and subordination (i.e., based on the structure of the individual clauses relative to one another or others). In the context of summarization, however, not all elements of a sentence are equally important. The most obvious solution to filtering out the non-essential elements is to include in the summary the most important (specific) clauses only. An example of this line of research is Marcu (1999), who uses rhetorical structure theory as a basis for determining discourse units, and uses the depth of the units in the discourse tree as a criterion for the extract-worthiness of the units. Although focusing on

single source documents, extensions for multi-document summarization seem straightforward. They include techniques relating to the frequency of the relevant units (using measures based on word-overlap to determine whether two discourse units are identical) combined with the depth of the different occurrences in the discourse trees.

- **Simplified-Sentence Extraction.**
  This approach is similar to discourse-unit extraction in the sense that the units of extraction are not necessarily whole sentences, yet the specifics are somewhat different. In particular, although it, too, removes non-essential clauses, this approach frequently also eliminates adverbial or prepositional modifiers because they are judged to be less important based on the frequency (or other any number of other criteria) of the words in the modifiers. By removing words from the sentence, the summarization system can select other sentences and thus increase the information density with respect to the most relevant information. There are generally two ways to carry out this extraction procedure: (1) The sentences are simplified in a pre-processing step, and original and simplified sentences are subjected to the sentence selection process (e.g., Conroy et al., 2005; Siddharthan et al., 2004); or (2) the sentences are simplified in a post-processing step based on the usefulness of modifiers and/or clauses (e.g., Daumé III and Marcu. (2005)).

- **Variable-Sized Sentence Extraction.**
  In contrast to the previous approaches, which are concerned with the removal of potentially superfluous clauses and/or modifiers, the approaches in this category *add* content to the extraction unit – as long as it is determined to be relevant. Note that this "opposing idea" does not necessarily result in the extraction of full clauses because not all information in the clauses is necessarily important. An example of an approach in this category is the system by Tucker and Sparck-Jones (2005), who create a graph over logical forms of a single document and then, based on multiple criteria, select sub-graphs for summarization. Unfortunately, they did not create full sentences, but only provided a list of surface representations of the sub-graphs. The approach proposed in this chapter also falls into this category.

In sum, the approaches described in this section illustrate the usefulness of particular sources of information for the summarization process. By combining the notions of

(word) frequency, connectivity (syntactic information, syntactic/semantic graphs), and the extraction of variable-sized units of information, one arrives at an approach that combines individual (singular) pieces of information into larger informational units and determines the units to be represented in the summary text based on their frequency (potentially along with additional importance criteria). The next section describes how I construct an approach in this spirit using techniques developed as part of the automatic Pyramid-style evaluation method presented in Chapters 3 and 4.

5.3 Summarization Based on Content Units

The multi-document summarization approach proposed in this section is rooted in an adaptation of the pyramid-generation algorithm of my automated evaluation method. To be precise, rather than extracting summary content units, I apply the algorithm to the original source documents to obtain content units. The content units, in turn, form the basis for determining the most important sentences and/or content units to be included in the ultimate summary text. Importantly, note that it is by no means guaranteed that the level of granularity that worked well on the reference summaries will also work well on the original documents. This is one of the issues to be investigated in the following sections.

In detail, the proposed summarization procedure translates the framework derived for the automatic evaluation system into the summarization context by way of the following four steps.

1. Determine the template instantiations in the original source documents;
2. cluster the template instantiations based on similarity;
3. combine the clusters based on co-occurrence statistics; and
4. select combined clusters based on attributes such as the number of elements in the combined clusters, their relative temporal ordering, and/or the relative structural ordering between combined clusters.

The first three steps in the summarization process are, in substance, identical to the evaluation scenario. The only difference is the nature of the underlying documents. From a practical perspective, however, it is not obvious that the settings are indeed

as similar as they seem. If they are, a straightforward implication would be that improvements in the evaluation scenario would directly carry over to the summarization scenario, which would clearly be beneficial.

The fourth step can be achieved in a number of different ways. The most promising approaches are the topic of the remainder of this section. The crudest approach would be to include the most frequent content units. The two main alternatives are the considerations of the temporal and structural orderings of the individual occurrences of the content units.

5.3.1 Frequency of Content Units

In its essence, the approach based on the most frequent content units (called FREQ in the experiments) corresponds to the direct application of the ideas underlying the automated evaluation method – the first step of creating a pyramid in particular – to the summary generation scenario. Recall that I use a bottom-up hierarchical clustering approach in order to determine the different content units and their relative importance as indicated by the number of members in each of the clusters representing the content units. The content units selected for the summary text are the informational units with the highest associated numbers of members in the cluster. If multiple informational units have the same frequency, one of them is selected randomly; owing to the nature of the pyramid, it is impossible to prefer specific informational units. Note, however, that the random selection is only relevant if the selection of all informational units with the appropriate frequency would exceed the length restriction.

Barring its emphasis on the variable size of the informational units, this approach would correspond to most frequency-based approaches to summarization. Yet, situations in which such an approach alone is not sufficient can easily be conceived, particularly in the context of multi-document summarization. The most obvious examples are occasions of documents sets that cover events over a certain period of time, implying that temporal relations are relevant for the extraction process, for instance, the espionage case between GM and VW in the DUC2005 dataset. In such a case, it should be clear that the conclusion of the affair (say, the settlement payment by VW to GM) should be included in the summary, no matter in how many documents those facts are mentioned. A frequency-based approach, however, cannot deduce that this information is any more relevant than other content units regarding the relevant parties. On the contrary, unless a disproportionate number of documents discuss the conclusion of

the case (only), this informational unit would receive very low frequency scores because newswire documents generally summarize some of the most important pieces of information at the beginning of new documents. As indicated, the inclusion of this information could be based on the temporal ordering of the information. To tackle this problem, the following section describes how temporal relations between content units can be used in the process of selecting relevant informational content for the summary.

5.3.2 Temporal Relations between Content Units

Besides highlighting the potential importance of accounting for temporal relations, the discussion at the end of the preceding section also suggests that the selection of informational content cannot *solely* be based on the temporal ordering of the informational units. For, in that case, only the latest information would be selected for the summary despite potentially relevant information in other documents. I therefore combine temporal ordering with the frequency of the informational units (called TEMP in the experiments). In particular, I consider the following three (partially overlapping) approaches, where X denotes the weight assigned to the temporal component(s) relative to the frequency ones:

1. **TEMP1-X.**
   The general formula for the importance of an informational unit according to this measure is given by $TEMP1 \cdot X + FREQ \cdot (1 - X)$, where $TEMP1$ denotes the relative position of the informational unit in the partial ordering generated based on the date of publication associated with the informational unit. If different instantiations are associated with different dates, the earliest date is employed for the generation of the partial ordering.

2. **TEMP2-X.**
   Paralleling TEMP1-X, the general formula for the importance of an informational unit in this case is given by $TEMP2 \cdot X + FREQ \cdot (1 - X)$, yet $TEMP2$ denotes the relative position of the informational unit in the partial ordering generated according to the earliest date in the same sentence as an instantiation associated with the combined cluster. If there is no such date occurrence, then the earliest date of publication is used (as in the case of TEMP1-X).

3. **TEMP-XN.**
   While TEMP1-X and TEMP2-X rely on the static relative importance of tempo-

ral relations and frequency, TEMP-XN follows a more variable route. In particular, the aforesaid example illustrating the usefulness of temporal relations highlighted the importance of selecting *some* informational units based on temporal relations; it did not suggest that *all* units should be selected. The TEMP-XN approach therefore decreases the importance of the temporal relations as additional informational units are selected; i.e., the first informational unit selected has the highest impact on temporal considerations, while the selection of subsequent units puts less and less weight on temporal considerations. The two main reasons for this course are: (1) if only one unit is selected then the most important unit is the one detailing the conclusion of the disagreement as opposed to the one detailing that there is a disagreement; and (2) it allows for a fixed maximum importance without regard to the number of informational units that are selected, because if the importance is increased in subsequent sentence selection steps, there is no upper bound for importance. Formally, the weight of the informational units available for selection as the N^{th} informational unit in the summary is given by $TEMP \cdot \frac{X}{N} + FREQ \cdot (1 - \frac{X}{N})$.

The first and second approach are quite similar. The difference is that one employs actual dates in the source document while the other exploits the publication dates of the source documents. The third approach is substantially different. It only assigns relatively high importance to the temporal orderings of the first (few) sentence(s) as opposed to asserting the same relative importance at all stages of the sentence selection process. Yet, although intuitive, it is clear that a focus on temporal relations is not the only possible way for improving information selection. The next section therefore explores an approach that utilizes the relative position of information in the source documents.

5.3.3 Structural Relations between Content Units

Another viable approach to determining the relevance of content units is their structural relation to each other (called STRUCT in the experiments). In other words, how do the positions of the content units in the source documents relate to one another? This extension to the frequency-based approach is rooted in the idea that the use of the position in which an informational unit occurs in the newswire documents is relevant to the information that should be selected for a summary. The intuition is that in newswire text, information that occurs at the beginning of a document tends to be

more important than information presented later on, particularly since newswire text frequently provides a brief synopsis or teaser at the outset of a document. A representative system exploiting this observation is the LEAD baseline (e.g., Barzilay and Lee, 2004). As before, I combine structural and frequency information, and investigate the following two approaches (with X defined as before):

1. **STRUCT-D-X.**
   The first approach only incorporates the position of a given sentence in the source *document* in which the informational unit occurs, called STRUCT-D. Since occurrences earlier in the (newswire) document are considered more important, the importance of the structural component is given by the inverse of the earliest position of the informational unit in a document, thereby assigning informational units in the first sentence a score of 1, while informational units in the last sentence have a score of $\frac{1}{numberofsentences}$. Formally, I consider the measure given by $\frac{1}{STRUCT-D} \cdot X + FREQ \cdot (1-X)$.

2. **STRUCT-P-X.**
   This approach is identical to STRUCT-D-X except for the fact that it considers the position of the sentence in the *paragraph*, termed STRUCT-P, as opposed to its position in the document; formally, the measure is given by $\frac{1}{STRUCT-P} \cdot X + FREQ \cdot (1-X)$.

In short, while the first approach is based on the same assumption as the LEAD baseline, i.e., information at the beginning of a source document is more important than information at later points in the document, the second approach assumes that the first sentences in a paragraph are more important.

Having discussed how informational units to occur in the summary can be selected, it is now necessary to consider how to construct the textual representation of the summary. The following section presents an approach to this problem. It is based on the importance of informational units in the source sentences.

5.4 Sentence Selection Based on the Importance of Informational Units

While a summary should ideally only contain exact informational units, in practice, this proves complicated for a multitude of reasons: (1) informational units are not nec-

essary full sentences; (2) incorrect syntactic analysis of the sentences in the original documents can cause grammatically incorrect output; and (3) exclusive use of informational units would result in pre-school phrasing owing to missing sub-clause structure and connectives between statements. For these reasons, I use sentences as the unit of extraction, which makes it necessary to obtain importance scores for the sentences in the documents based on the informational units they contain. To this end, I assign each sentence in the original documents a weight based on the informational units in the sentences already selected for the summary, the informational units in the sentences, and the number of words in the sentence.

Note that information about the number of words in a sentence is important because without taking the length of the sentence into account, the approach would prefer longer sentences over shorter ones. The length of the sentence is therefore chosen to normalize the score. Informational content in the sentences already selected for the summary is important because the aim is to have as many relevant informational units in the summary as possible (given the length requirements). To achieve this, it is necessary to penalize multiple selections of the same informational content.

Formally, the score of a sentence is given by

$$\frac{\sum_{a=1}^{n} w_a}{wordcount},$$

where n denotes the number of informational units in the sentence that did not occur in the sentences already selected for the summary, w_a represents the weight of the a^{th} informational unit in the sentence that does not occur in the sentences already selected for the summary, and *wordcount* signifies the number of words in the sentence.

5.5 Maximum Marginal Relevance in a Pyramid-Based Summarization Process

The last major issue with regard to the application of the concepts from my automated Pyramid-style evaluation method to automatic summarization in need of investigation is the marginal relevance of the different content units in the sentence selection process. "Marginal Relevance" (MR) is a measure of the importance of a sentence given the previously selected sentence. For example, if all of the informational content of a given sentence is contained in the sentences that have already been selected for the summary, its marginal relevance is low. It should consequently not be included in the

summary.

In methods based on word frequency, a prominent issue is the selection of different sentences containing the same content, i.e., sentences with a low *maximum* marginal relevance given that some other sentence is already part of the summary. Carbonell and Goldstein (1998) proposed the concept of "Maximum Marginal Relevance" (MMR) as a solution to the problem of repetitive information. In its essence, MMR selects sentences with the highest weighted average of the importance score and a similarity score between the sentence being considered for inclusion and those already contained in the summary. Sentences to be included according to this approach should have high importance and low similarity to other sentences in the summary.

In this context, from a theoretical perspective, the approach presented in this chapter should cluster similar content together using the techniques from my automated evaluation scheme, whereupon the sentence selection process selects the clusters to be included in the summary. If a cluster were represented twice in the summary, it would be penalized as it "wastes" space. As a result of this sequence of processing steps, it should not be necessary to use an MMR-style sentence selection process (since its purpose is already explicitly being accounted for). However, in view of the significant differences between my syntactically motivated multi-word unit approach and (simple) word overlap approaches, I will also investigate whether the use of additional MMR-style sentence selection has any impact on the quality of the informational content of the summary. The following section explores the issues raised in this section from a practical perspective.

5.6 Experiments

Before delving into the experiments used for the evaluation of the summarization approach presented in this chapter, the following provides a short summation of the proposed approach. In particular, I use the clustering and template techniques developed for my automatic evaluation approach to create a pyramid containing variable-sized informational units along with their frequencies in the source documents of the information presented in the source documents. Subsequently, sentences are selected based on the frequency, temporal and/or structural information of the content units that occur within the sentences, potentially using MMR in order to remove sentences that are too similar. The outcome is a summary comprising the sentences containing the highest-scored informational units from the source documents.

To evaluate the various features of this new approach to multi-document summarization, I use ROUGE (Lin and Hovy, 2003) and the automated evaluation methods developed in Chapters 3 and 4. Even though my automated evaluation method has been shown to correlate better with the (original) manual Pyramid method than does ROUGE, I nonetheless report both schemes, as my summarization and evaluation methods use the same techniques. By reporting both measures, I avoid the criticism that the new summarization approach is biased towards the evaluation method. In addition, this step improves the comparability of the results with previous work.

In terms of datasets, I once again use the 10-cluster subset of the DUC2005 dataset for the initial exploratory experiments and the remaining 40 clusters for the final evaluation of the summarization system.

5.6.1 Experiment 1: General Pyramid Statistics and the Influence of Different System Settings

The objective of this experiment is to explore a number of aspects relating to the use of the techniques underlying my evaluation method to automatically generate summaries. The first step in this regard is the adaptation of the granularity and specificity of the clusters generated using the relevant techniques. At this stage, the points of interest are the optimal number of clusters for the summarization of newswire documents and the average and maximum number of elements in the clusters. This information provides a basis for the subsequent experiments, and enables one to draw conclusions regarding the impact of cluster size on different aspects of the system. In particular, I consider the impact of different decisions with respect to template similarity (ts) and cluster overlap (co) on the statistics for the resulting pyramid.

Template similarity varies the similarity required in order to achieve a positive match between two template instantiations. It is determined using the similarity of the corresponding constituents of the relevant templates using the manually determined associations between the constituents of the templates. The cluster overlap setting is the threshold of co-occurrence between two individual clusters such that two clusters are combined (i.e., the percentage of times that the two clusters occur within the same sentence). The intervals for the settings reported were determined in initial experiments that selected the most promising range for good results.

Table 5.1 presents the results of explaining the impact of different settings for the pyramid generation process on the number of clusters and their relative sizes. They

indicate that the settings significantly influence size and number of clusters. In the case of template similarity, different settings tend to increase the number of clusters by 20% between the ts= 0.5 and ts= 0.7 settings. Likewise, cluster overlap results in a difference of about 30% in the number of clusters for the different settings in the range being investigated. Similar differences hold true for the average number of units in the clusters. The maximum number of units in the cluster, on the other hand, is virtually identical regardless of the setting being explored.

Method	Avg. # of clusters	Average	Maximum
original (ts = 0.55 & co = 0.65)	150	2.2	6
template similarity (ts) = 0.5	141	2.6	7
template similarity = 0.6	153	2.2	6
template similarity = 0.7	170	1.8	6
cluster overlap (co) = 0.5	120	2.6	6
cluster overlap = 0.6	126	2.3	6
cluster overlap = 0.7	135	2.3	6
cluster overlap = 0.8	160	2.0	6

Table 5.1: Results of Experiment 1. The impact of different settings for the automatic pyramid generation on the number of clusters and their relative size. The table shows the statistics for a number of different settings for determining the size and composition of the content units, i.e., when two template instantiations are considered the same (ts) and when the co-occurrence between two clusters is sufficient (co). "Original" denotes the settings for template similarity and cluster overlap that were determined in the preceding chapter for the evaluation of summarization systems.

5.6.2 Experiment 2: Optimal System for Frequency-Based Pyramid Summarization

Having obtained an overview of approximate cluster sizes and statistics relating to different settings of the system, I turn to the impact of the different settings for template similarity and cluster overlap on the resulting summaries. The results of this investigation are depicted in Table 5.2. They suggest that a template similarity setting of 0.5 and a cluster overlap of 0.6 achieve the highest evaluation scores. On account of these

findings, the remaining experiments involve two setting combinations only, namely, ts= 0.5 & co= 0.6, and ts= 0.6 & co= 0.6.

PGM	**SGM**	**ROUGE-SU4**	**AP**	**PAP**
ts=0.5 & co=0.5	FREQ	0.12	0.16	0.17
ts=0.5 & co=0.6	FREQ	**0.13**	**0.17**	**0.18**
ts=0.5 & co=0.7	FREQ	0.12	**0.17**	0.17
ts=0.6 & co=0.5	FREQ	0.09	0.15	**0.17**
ts=0.6 & co=0.6	FREQ	**0.12**	**0.16**	**0.17**
ts=0.6 & co=0.7	FREQ	0.10	0.15	0.16
Low DUC2005		0.02	0.06	0.07
High DUC2005		0.12	0.18	0.19

Table 5.2: Results of Experiment 2. The impact of different settings for template similarity and cluster overlap on the summarization process. The table shows the evaluation results using a number of (semi-)automatic evaluation methods for different settings of the pyramid generation. "PGM" denotes the automatic pyramid generation method, i.e, the settings to create the pyramid, "SGM" signifies the summary generation method, i.e., the method for selecting content units to be included the summary, "AP" represents the automated Pyramid-style evaluation method presented in Chapter 4, and "PAP" the partially automated Pyramid-style evaluation method developed in Chapter 3. As before, "co" denotes cluster overlap and "ts" template similarity. The highlighted numbers constitute the highest results. "Low DUC" and "High DUC" denote the worst and best summarization systems in the DUC2005 evaluation, respectively.

5.6.3 Experiment 3: The Impact of Temporal and Structural Relations on Pyramid Summarization

Although simple frequency-based sentence selection provides satisfactory results, in view of its aforementioned shortcomings (Section 5.3.1), this experiment explores the impact of using temporal and structural relations on sentence selection.

The results of the investigation, presented in Tables 5.3 and 5.4, reveal that the inclusion of temporal and structural relations improves the results, as may have been expected based on the discussion in the relevant sections. However, there is a marked

PGM	SGM	ROUGE-SU4	AP	PAP
ts=0.5 & co=0.6	TEMP1-0.1	0.13	0.17	0.18
ts=0.6 & co=0.6	TEMP1-0.1	0.12	0.16	0.17
ts=0.5 & co=0.6	TEMP1-0.2	0.12	0.17	0.18
ts=0.6 & co=0.6	TEMP1-0.2	0.12	0.16	0.17
ts=0.5 & co=0.6	TEMP1-0.3	0.12	0.16	0.17
ts=0.6 & co=0.6	TEMP1-0.3	0.11	0.16	0.17
ts=0.5 & co=0.6	TEMP2-0.1	0.13	0.17	0.18
ts=0.6 & co=0.6	TEMP2-0.1	0.12	0.16	0.17
ts=0.5 & co=0.6	TEMP2-0.2	0.14	0.18	0.19
ts=0.6 & co=0.6	TEMP2-0.2	0.14	0.17	0.18
ts=0.5 & co=0.6	TEMP2-0.3	0.13	0.16	0.18
ts=0.6 & co=0.6	TEMP2-0.3	0.13	0.16	0.17
ts=0.5 & co=0.6	TEMP-0.1N	0.14	0.18	0.18
ts=0.6 & co=0.6	TEMP-0.1N	0.14	0.18	0.18
ts=0.5 & co=0.6	TEMP-0.2N	**0.15**	**0.19**	**0.19**
ts=0.6 & co=0.6	TEMP-0.2N	0.15	0.18	0.18
ts=0.5 & co=0.6	TEMP-0.3N	0.14	0.18	0.18
ts=0.6 & co=0.6	TEMP-0.3N	0.14	0.18	0.17

Table 5.3: Results of Experiment 3 (Part 1). The impact of the use of temporal relations on informational content. For a decoding of the abbreviations, refer to Table 5.2.

difference in the scale of improvement achieved by the two measures. The enhancement attained by accounting for temporal relations is significantly higher than that accomplished with structural relations. By far the greatest improvement is obtained when using the "TEMP-0.2N" configuration, which selects the first few sentences with significant weight on temporal relations, while later sentences are solely selected based on the frequency of the informational units.

PGM	SGM	ROUGE-SU4	AP	PAP
ts=0.5 & co=0.6	STRUCT1-0.1	0.13	0.17	0.18
ts=0.6 & co=0.6	STRUCT1-0.1	0.12	0.16	0.17
ts=0.5 & co=0.6	STRUCT1-0.2	**0.13**	**0.18**	**0.18**
ts=0.6 & co=0.6	STRUCT1-0.2	0.13	0.17	0.18
ts=0.5 & co=0.6	STRUCT1-0.3	0.12	0.17	0.17
ts=0.6 & co=0.6	STRUCT1-0.3	0.12	0.16	0.17
ts=0.5 & co=0.6	STRUCT2-0.1	0.13	0.17	0.18
ts=0.6 & co=0.6	STRUCT2-0.1	0.12	0.16	0.17
ts=0.5 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17
ts=0.6 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17
ts=0.5 & co=0.6	STRUCT2-0.3	0.11	0.15	0.16
ts=0.6 & co=0.6	STRUCT2-0.3	0.12	0.15	0.17

Table 5.4: Results of Experiment 3 (Part 2). The impact of the use of structural relations on informational content. For a decoding of the abbreviations, refer to Table 5.2.

5.6.4 Experiment 4: The Impact of MMR on Pyramid Summarization

Last but not least, I investigate the usefulness of MMR in the summary generation setting. For completeness, I explore its impact on a number of different methods that provided good results in the preceding experiments (namely,"TEMP2-0.2," "TEMP-0.2N," and "STRUCT1-0.2"). Table 5.5 shows that the use of MMR in the sentence selection process only has a slight influence on the evaluation scores. It is negligible because it only makes a difference in two cases ("ts=0.6 & co=0.6, TEMP2-0.2" and "ts=0.5 & co=0.6, TEMP2-0.2") – in the former, in a positive, and in the latter, in a negative direction. Since my MMR approach is based on word overlap as opposed

to syntactic similarities, this would seem to indicate that my pyramid-based summarization approach correctly identifies the major similarities between sentences without requiring further removal of similar information based on word overlap information.

MMR	PGM	SGM	ROUGE-SU4	AP	PAP
No MMR	ts=0.5 & co=0.6	TEMP2-0.2	0.14	0.18	0.19
No MMR	ts=0.6 & co=0.6	TEMP2-0.2	0.14	0.17	0.18
No MMR	ts=0.5 & co=0.6	TEMP-0.2N	0.15	0.19	0.19
No MMR	ts=0.6 & co=0.6	TEMP-0.2N	0.15	0.18	0.18
No MMR	ts=0.5 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17
No MMR	ts=0.6 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17
MMR	ts=0.5 & co=0.6	TEMP2-0.2	0.14	0.18	**0.18**
MMR	ts=0.6 & co=0.6	TEMP2-0.2	0.14	**0.18**	0.18
MMR	ts=0.5 & co=0.6	TEMP-0.2N	0.15	0.19	0.19
MMR	ts=0.6 & co=0.6	TEMP-0.2N	0.15	0.18	0.18
MMR	ts=0.5 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17
MMR	ts=0.6 & co=0.6	STRUCT2-0.2	0.12	0.16	0.17

Table 5.5: Results of Experiment 4. The impact of MMR on the summarization process, i.e., with respect to the number of clusters and their sizes. For a decoding of the abbreviations, refer to Table 5.2.

5.6.5 Experiment 5: Overall Performance of Pyramid Summarization in Relation to other Systems

The experiments up to this point have investigated the impact of a number of individual features considered for the adaptation of the automated Pyramid-style evaluation process to the summarization task. Table 5.6 summarizes the results when comparing the (best) ensuing system – "No MMR, ts=0.5 & co=0.6, TEMP-0.2N" – to other state-of-the-art summarization methods. They show that it outperforms the best system in the DUC2005 competition based on the ROUGE-SU4 evaluation, while it performs as well as the best DUC2005 system when evaluated via my partially automated and fully automated evaluation methods.

Method	ROUGE-SU4	AP	PAP
Pyramid-Based Summarization	**0.15**	**0.19**	**0.19**
Low DUC2005	0.02	0.06	0.07
High DUC2005	0.12	0.18	**0.19**

Table 5.6: Results of Experiment 5. The overall performance of the proposed Pyramid-based summarization approach compared to other state-of-the-art multi-document summarization systems. For a decoding of the abbreviations, refer to Table 5.2.

5.7 Remarks Relating to Statistical Significance and Confidence Intervals

Paralleling the previous chapter, while statistical significance and confidence intervals can in principle be computed, the results tend to be rather inconclusive. Passonneau et al. (2005) report confidence intervals on the Pyramid scores of approximately 0.1 units. Given that Pyramid scores tend to be quite similar, minor changes within the confidence intervals of a given system's Pyramid scores could (already) influence the ranking of the systems.

Passonneau et al. (2005) furthermore report the results of several Wilcoxon rank sum tests ($\alpha = 0.5$). They suggest that the highest number of statistically significant differences between a system and its competitors is 8. Considering that they look at a total of 30 runs, this result indicates that there are no significant differences between most systems. The situation becomes even more "depressing" when using a more conservative test, Tukey's honest significant difference method, which avoids the Type-1 error that might result from the combinatorial use of significance tests. In Passonneau et al. (2005)'s explorations, this test shows a single system to be – at best – better than two other systems, meaning once more that there are hardly any significant differences. In other words, the use of statistical significance testing does not provide much useful information, which is the reason the present work does not consider these tools.

5.8 Examples of the Information Selected by the Pyramid Summarization System

The experiments in the preceding section investigated the impact of a variety of aspects on the performance of the pyramid generation process and compared the approach presented in this chapter to a number of other systems. They did not, however, illustrate the effects of the pyramid generation algorithm on the content selection procedure. To this end, in this section, I provide examples demonstrating the workings of the general process, the ensuing content units, and the summaries obtained when applying the process to a collection of documents.[3] As in the preceding chapters, they are taken from the document sets for the initial summaries of the TAC2009 "Update Summarization" track. Recall that the document sets contain only 10 documents (each), which makes them prime resources on which to illustrate the workings of the algorithm.

The first example, presented in Figure 5.2, is from a document set about the relations between India and Pakistan (cf. reference summary). The objective is to illustrate the generally good performance while at the same time highlighting some problems. In this regard, note that the representative sentences from the source documents resemble each other in that they contain three similar facts: (1) there was a troop withdrawal, (2) India withdrew troops, and (3) the withdrawal happened in/around Kashmir.

Although the first fact is represented in all six sample sentences, the proposed summarization system does not recognize the similarity between "reduction" and "withdrawal." The content unit relating to the withdrawal (Fact 1) therefore only has five contributors. Likewise, the content unit pertaining to the second fact, too, has five contributors. Unfortunately, the content unit concerning the third fact only has two contributors because "from the region" is not correctly associated with "in Kashmir." For simplicity, the other potential content units, e.g., that Singh made a trip lasting two days, are omitted from Figure 5.2. Owing to the perfect correlation between the first and second content unit, they are combined as part of the clustering algorithm developed for the fully automatic evaluation scheme, creating the (final) content unit that "Indian troops were withdrawn." Extrapolating this process to all sentences in the document set and mapping the content units to sentences and then sentence scores gives rise to the (automatically generated) summary presented in Figure 5.3.[4] Observe that

[3] Note that I do not account for the topic statement for the summaries, i.e., the examples involve the generation of generic summaries.

[4] For transparency, the brackets following each of the summary sentences indicate the relevant content units. The order of sentences in the figures containing the summary represent the order of selection

Reference Summary Created by Annotator A (100 words):

Since they became separate nations in 1947, India and Pakistan have fought two wars over Kashmir, the Himalayan province which was split between them. Kashmir is India's only majority Islam province, and an Islam insurgency began on the Indian side in 1989. More than 45,000 people were killed. A bilateral cease fire was pronounced in 2003, and Prime Minister Singh of India reduced troop levels by 40,000 in a force of 500,000. Peace talks began in 2004 and in March 2005, bus service between the two countries began, the first tangible result of the peace talks.

Sentences from the Source Documents Containing Similar Information:

(1) Indian Prime Minister Manmohan Singh, who arrives here on a two- day visit next week, announced on Thursday that orders had been given for the reduction of troops inside Kashmir.

(2) Singh's trip to Kashmir came as India began withdrawing some of its troops from the region, a goodwill gesture to war-weary Kashmiris and Pakistan ahead of planned talks between the nuclear armed rivals on the problem due next month.

(3) The proposals have been greeted with little enthusiasm by India, although it has since announced its troop withdrawal.

(4) Singh's two-day trip came as India began withdrawing some of its troops on the border, a goodwill gesture to war-weary Kashmiris and rival Pakistan.

(5) India has cited a decline in separatist violence as the main reason for its troop withdrawal, which reportedly will be about 40,000 of the half-million stationed in Kashmir.

(6) On Wednesday, Singh, the Indian leader, expressed a commitment to make peace with Pakistan, while Islamabad hailed India's withdrawal of some of India's troops in Kashmir — a move hoped to spur the peace process.

Content Unit (Fact 1 and 2) Extracted From the Sample Sentences:

withdrawing some of its troops
its troop withdrawal
withdrawing some of its troops
its troop withdrawal
India's withdrawal of some of India's troops

Figure 5.2: Example 1 (System Input). A human reference summary for document set D0901A-A (TAC2009), a sample of similar sentences from the document set, and representative content units extracted from the sentences.

it contains the following pyramid SCUs: “India reduced troop levels” (weight 4 in the pyramid created based on the manual reference summaries), “bus service between the two countries began” (weight 4 in the pyramid), “India rejected redrawing the border” (weight 2 in the pyramid), “Kashmir is divided between India and Pakistan” (weight 1 in the pyramid), “reduced troops was a goodwill gesture” (weight 1 in the pyramid). Overall, the example shows that the content unit approach works well, but does have a few issues.

```
Pakistan and India agreed in February to re-establish a bus link between the Indian-
 and Pakistan-controlled portions of Kashmir as part of moves to improve relations
 between the two longtime rivals. [(31 words) content units: establish bus link;
 improve relations; Pakistan and India agreed]

Singh's two-day trip came as India began withdrawing some of its troops on the
 border, a goodwill gesture to war-weary Kashmiris and rival Pakistan. [(24 words)
 content units: Singh's trip; withdrawing some of its troops; goodwill gesture to
 Pakistan]

Kashmir is divided between India and Pakistan but both claim the region in its
 entirety.[(15 words) content units: Kashmir is divided; claim the region]

A third round of talks with India's government can't start until it agrees to let
 Kashmiri leaders visit Pakistan to discuss peace proposals with its government and
 guerrilla commanders in [the part of Kashmir under Pakistan's control, Farooq
 said]. [because of length sentence truncated starting at the part; (30 words)
 content units: talks with India's government; discuss peace proposals; let leaders
 visit]
```

Figure 5.3: Example 1 (Summary). The summary for document set D0901A-A (TAC2009) generated by my Pyramid Summarization System.

The second example, illustrated in Figure 5.4, is taken from a document set on the development of the situation in Northern Ireland following the suspension of the Good Friday peace accord (*cf.* reference summary). Note that the representative sentences from the source document are quite a bit more complex than those in the preceding example. Correspondingly, the system needs to cope with issues regarding incorrect syntactic structure and modifier attachment, which will be pointed out at the relevant points. The point of the example is the identification of the content unit(s) regarding the “end of violence.”

From a naïve perspective, all sentences clearly contain this fact. Problematically,

of the sentences, i.e., the first sentence in each figure is the first sentence selected and therefore the most important according to the selection algorithm

Reference Summary Created by Annotator B (100 words):

The 1998 Good Friday peace accord was suspended in 2002 due to violent activity by the Irish Republican Army (IRA). An appeal in April 2005 by Sinn Fein leader Adams to the IRA to abandon armed struggle was endorsed in May by Irish Prime Minister Ahern. All interested parties awaited a response from the IRA. In late June the U.S. State Department supported British Prime Minister Blair's and Ahern's call for a response. In late July there was a first glimmer of hope as three senior Sinn Fein leaders resigned from the ruling body of IRA's military wing.

Sentences from the Source Documents Containing Similar Information:

(1) Britain expects an "imminent" statement from the IRA as to whether the [...] group is to agree to an appeal from its political wing to abandon violence [...].
(2) London [...] are awaiting for an Irish Republican Army response to a call from Gerry Adams, leader of the group's Sinn Fein political wing, for an end to violence.
(3) In April, Adams called on the IRA to "fully embrace and accept" democratic means and end all paramilitary activities.
(4) "To move forward, we need a clear, unambiguous end of all paramilitary and criminal activity and we need to see [...] (weapons) decommissioning," Ahern said.
(5) But he said the pact's key goal [...] would be revived only "in the context of a complete end to IRA paramilitarism and criminality [...]."
(6) Last month, Sinn Fein leader Gerry Adams, a reputed IRA commander, appealed to IRA members to leave behind their "armed struggle" in favor of democratic politics.
(7) Protestant factions in Northern Ireland are adamant that there can be no political progress toward a lasting peace settlement without a move by the IRA to end all paramilitary and criminal activity [...].
(8) In April, Since Sinn Fein leader Gerry Adams made a direct appeal to IRA "volunteers" to give up all violence and adopt democratic methods, since then the Catholic paramilitary group has been holding meetings to discuss a response.
(9) The United States, he said, welcomed statements made this week by [...] Tony Blair and [...] Bertie Ahern that the key to moving forward was for the Irish Republican Army (IRA) [...] to end all violence unequivocally.
(10) Protestant factions in Northern Ireland are adamant that there can be no political progress toward a lasting peace settlement without a move by the IRA to end all criminal and paramilitary activity.

Content Units Extracted from the Sample Sentences:

Contributors of Content Unit 1:
- to abandon violence
- to give up all violence

Contributors of Content Unit 2:
- for an end to violence
- to end all violence unequivocally

Contributors of Content Unit 3:
- end all paramilitary activities
- end of all paramilitary and criminal activity
- a complete end to IRA paramilitarism and criminality
- IRA [...] to end all paramilitary and criminal activity
- IRA to end all criminal and paramilitary activity

Figure 5.4: Example 2 (System Input). A human reference summary for document set D0909B-A (TAC2009), a sample of similar sentences from the document set, and content units extracted from the sentences.

```
The 1998 Good Friday peace deal paved the way for a Protestant-Catholic
 power-sharing assembly, but that was suspended more than two years ago amid
 allegations of IRA espionage. [(28 words) content units: Good Friday peace deal in
 1998; paved for assembly; suspended two years ago]

Ahern said he was hopeful that the IRA would issue a declaration within the next few
 months spelling out whether it will disarm and cease all threatening activities.
 [(28 words) content units: IRA issue declaration; it will cease all threatening
 activities]

Protestant factions in Northern Ireland are adamant that there can be no political
 progress toward a lasting peace settlement without a move by the IRA to end all
 criminal and paramilitary activity. [(32 words) content units: progress toward
 peace; IRA end paramilitary activity]

The IRA declared a ceasefire in its campaign to end British rule [in Northern
 Ireland before the 1998 peace deal, which largely ended 30 years of sectarian
 violence in the province]. [because of length sentence truncated starting at Hatch;
 (12 words) content units: IRA declared ceasefire]
```

Figure 5.5: Example 2 (Summary). The summary for document set D0909B-A (TAC2009) generated by my Pyramid Summarization System.

the wording varies considerably. For example, "paramilitary activity" and "violence" imply the same thing, but WordNet does not allow this inference. For this reason, the relevant parts of the second ("for an end to violence") and ninth ("to end all violence unequivocally") sentences are clustered together, while the same is true for the first ("to abandon violence") and eighth ("to give up all violence") sentences, and the third ("end all paramilitary activities"), fourth ("end of all paramilitary and criminal activity"), fifth ("a complete end to IRA paramilitarism and criminality"), seventh ("to end all paramilitary and criminal activity") and tenth ("to end all criminal and paramilitary activity") sentences. The sixth sentence cannot be associated with any of these clusters. Incorrect syntactic attachment of the fact that the IRA should end violence in the third, eighth and ninth sentences prevents the identification of the similarity between "paramilitary activity" and "violence." As a result, the three clusters remain separated. For the "paramiltary" clusters, in turn, sufficient evidence is found that the IRA should give up those activities for them to be joined together. The three resulting clusters for the single actual unit of information are provided in the figure.

The summary generated for this document set is shown in Figure 5.5. The content of the summary is acceptable in terms of the pyramid SCUs it contains, though not as good as the summary for the previous example: "Good Friday pact was agreed in

1998" (weight 4 in the pyramid), "Good Friday pact was a peace pact" (weight 4 in the pyramid), "Good Friday pact set up a joint Catholic-Protestant administration in Northern Ireland" (weight 3 in the pyramid), "Good Friday pact suspended in 2002" (weight 3 in the pyramid), "All interested parties awaited a response from the IRA" (weight 1 in the pyramid). In essence, this example shows problems of the summarization approach that are based on both syntactic and semantic information.

The final example, presented in Figure 5.6, illustrates an instance in which a lot of sentences contain the same information. It is taken from a document set on the retirement of a Supreme Court Justice and the appointment of her successor (*cf.* reference summary). This example concentrates on the content unit(s) relating to the announcement of Ms. O'Connor's resignation.

With the limited anaphora resolution performed in the context of this thesis, the pronoun "she" is resolved to "Sandra Day O'Connor" and, as such, the fact that O'Connor retires is relatively easy to determine in most sentences (since "retire" and "resign" are closely related in WordNet). Unfortunately, Sentence 8 is parsed incorrectly and thus attaches "retirement" to "historic chance" as head of the noun phrase, which clearly results in a missed identification; Sentence 8 is not identified correctly because the fact the O'Connor is retiring is 'hidden' in noun phrases not directly related to the retirement. All eight other instances are clustered together correctly.

Note that since WordNet considers "say," "announce," and "explain" very closely (with respect to at least one of their word senses), the fact that "O'Connor said" something is expressed in all but Sentence 4. The result is that the clusters containing those two facts are joined together because of the high overlap between the two clusters (even though the fact that she did the announcement herself is not of primary importance). Yet, none of the other facts co-occur sufficiently with this cluster to be combined – apart from the information that O'Connor is a justice. As this information also occurs frequently in sentences that do not deal with the retirement, it is not joined with the content unit. The contributors of the content unit relating to Ms. O'Connor's resignation/retirement are presented at the bottom in the figure.

Reference Summary Created by Annotator B (100 words):

Sandra Day O'Connor, the first woman to serve on the US Supreme Court, is retiring. She gave no reason but a Court spokesman indicated her need to spend more time with her husband who has Alzheimer's. O'Connor, the first woman to serve on the Court was considered a conservative when seated but became more moderate over the years. President Reagan appointed her in 1981. President Bush now has his first opportunity to nominate someone to the high court and promises to do so in a timely manner. Liberals warned against picking an "extremist" while Republicans argued for a strict conservative.

Sentences from the Source Documents Containing Similar Information:

(1) Sandra Day O'Connor, the first woman ever appointed to the US Supreme Court, said Friday that she is retiring, giving US president George W. Bush his first opportunity to appoint a justice.

(2) "This is to inform you of my decision to retire from my position as an associate justice of the Supreme Court of the United States, effective upon the nomination and confirmation of my successor," she said in a letter to Bush.

(3) O'Connor, 75, did not explain why she was resigning.

(4) In a statement confirming O'Connor's resignation, Bush said he will pick her successor in a timely manner so her vacancy can be filled by the time the court resumes work in October.

(5) As President Bush searches for someone to succeed Justice Sandra Day O'Connor, who announced her resignation Friday, the experiences with Warren [...] provide reminders that justices often don't vote the way presidents expect.

(6) On Sunday, a seminar will focus on who will fill the Supreme Court seat that will be vacated by Justice Sandra Day O'Connor, who announced her retirement July 1.

(7) Republicans and Democrats are geared up for a major political fight over a successor to Supreme Court Justice Sandra Day O'Connor, who announced her retirement last Friday.

(8) Conservatives [...] see in the retirement of moderate Justice Sandra Day O'Connor a historic chance to reshape the nine–member high court, and fulfill a fervent wish to overturn the [...] case which guaranteed women the right to an abortion.

(9) US Republicans and Democrats geared up Saturday for a major political battle over a successor to Supreme Court Justice Sandra Day O'Connor, who has announced her retirement.

(10) Sandra Day O'Connor, the first woman appointed to the US Supreme Court and a frequent swing vote, announced her retirement Friday, setting up a fierce political showdown for her seat.

Content Unit (on Resignation/Retirement) Extracted from the Sample Sentences:

Sandra Day O'Connor [...] said [...] she is retiring
O'Connor did not explain [...] she was resigning
O'Connor's resignation
Sandra Day O'Connor, who announced her resignation
Sandra Day O'Connor, who announced her retirement
Sandra Day O'Connor, who announced her retirement
the retirement of Sandra Day O'Connor
Sandra Day O'Connor [...] announced her retirement

Figure 5.6: Example 3 (System Input). A human reference summary for document set D0936G-A (TAC2009), a sample of similar sentences from the document set, and the content unit extracted from the sentences.

The summary generated by the Pyramid-style approach presented in this chapter is presented in Figure 5.7. The content selected for the summary is quite good. It correctly identifies that "O'Connor is a Supreme Court Justice" (weight 4 in the pyramid), "O'Connor announced her retirement from Supreme Court" (weight 4 in the pyramid), "President Bush will name a successor" (weight 3 in the pyramid), "O'Connor was the first female on the Supreme Court" (weight 2 in the pyramid), "Potential nominees for the position have been mentioned" (weight 2 in the pyramid) and "Political leaders and lobby groups prepared for battle over the nomination" (weight 1 in the pyramid). The only information contained in the summary that was not part of the pyramid (created based on the manual reference summaries) is the President's consultation with the Senate, and the duplicate information that there is going to be a fierce battle for the seat. In sum, this example highlights potential problems relating to the inclusion of extraneous information for inclusion into the content units.

```
US President George W. Bush on Wednesday chastised critics of Attorney General
 Alberto Gonzales as a potential candidate for the recently-vacated seat in the
 Supreme Court. [(26 words) content units: US President Bush; critics of Gonzales;
 Attorney General Gonzales; candidate for Supreme court; vacated seat]

Sandra Day O'Connor, the first woman appointed to the US Supreme Court and a
 frequent swing vote, announced her retirement Friday, setting up a fierce political
 showdown for her seat. [(30 words) content units: first woman on supreme court;
 O'Connor on supreme court, announced her retirement]

He also promised to "be deliberate and thorough" in making the choice and will
 consult with the Senate which will confirm his nominee. [(23 words) content units:
 will consult with Senate; confirm his nominee]

"He said he's going to pick a strong conservative, and I think that's going to cause
 a battle no matter what," [Hatch told ABC television Tuesday]. [because of length
 sentence truncated starting at Hatch; (21 words) content units: pick conservative;
 cause a battle]
```

Figure 5.7: Example 3 (Summary). The summary for document set D0936G-A (TAC2009) generated by my Pyramid Summarization System.

Overall, these examples show that the summarization approach developed in this chapter successfully identifies a lot of relevant information, particularly for the earlier selected sentences. The quality of the sentences selected at later stages decreases as is illustrated by the fact that they generally contain fewer high-ranking SCUs. The first and second sentence typically contain high-weight SCUs while the later sentences

mostly contain those only found in one or at most two of the reference summaries.

5.9 Discussion

This chapter investigated the applicability of the ideas underlying the automatic evaluation method developed in Chapters 3 and 4 in the context of the generation of summaries from multiple source documents. The working hypothesis was that the relevant Pyramid-style approach, which is based on variable-sized informational units with specific syntactic relations between constituents as well as associated frequency information, should also provide good results when employed for summary generation. Indeed, this basic approach already yields good performance, achieving similar results as the best systems in the DUC2005 competition. By incorporating temporal relations into the selection process for some of the sentences to be included in the summary text, I was able to boost performance such that the ensuing system either outperforms or equals the performance of all of the systems that participated in the DUC2005 competition.

It should, however, be noted that the state-of-the-art in automatic summarization has evolved since 2005. While a number of recent efforts have used the DUC 2005 dataset (e.g., He et al. (2008); Ma et al. (2008); Ouyang et al. (2007); Nastase (2008); Chowdary and Kumar (2009); Bosma (2009)), and the system developed in this chapter performs well compared to these systems, most of the latest summarization approaches are evaluated on more recent datasets. As a result, this system cannot claim to be the best system "out there." This chapter nonetheless shows that the evaluation approach based on informational content units translates well into summary generation.

It is my belief that further research in the ultimate selection process of the content units which should occur in the summary would result in a very competitive state-of-the-art summarization system. For example, information need and prior user knowledge are not exploited. Furthermore, relationships between information units are not considered, e.g., in the form of broad topics. A useful side-effect of the experiments presented in this chapter is that there does not (per se) seem to be a bias of the Pyramid-based summarization method towards my automated evaluation method.

Chapter 6

Domain-Independent Shallow Sentence Ordering

6.1 Introduction

Up to this point, this thesis has investigated automatic summarization and its evaluation from the perspective of identifying and creating summaries with high-quality informational content. Some of the issues being explored were the (optimal) size of a content unit, whether the evaluation of informational content can be automated, and whether the techniques used for the evaluation can be used in order to select informational content to generate summaries. The preceding chapter, in particular, considered the extraction of sentences with the highest scored informational content as the sentences that should be included in a summary. However, in order to create an informative summary, besides retrieving the relevant information, the information has to be presented in such a way that it makes sense and is readily accessible to the reader. This chapter, therefore, explores the following question: How good is the ordering of the information in a summary?

A wide variety of issues have an impact on this seemingly simple query. In the context of determining the best possible structure for a summary, they range from such fundamental concerns as how to measure the quality of a given structure and establishing the necessary processing steps in order to create a structurally sound summary, to more practical aspects such as whether structural considerations should be a post-processing step that orders the informational content in the best possible way, or whether these attributes should (already) be taken into account when selecting the informational content. An in-depth study of all of these issues is beyond the scope of this thesis; however, I make a start by focusing on how one can measure the quality of the ordering of information in system summaries. In particular, I consider the following question: given informational content in the form of individual sentences, which ordering of the sentences yields the best summary from a structural perspective?

Sentence ordering is a problem in many natural language processing tasks. While it was, historically, mainly considered a challenging problem in (concept-to-text) language generation tasks (Reiter and Dale, 2000), more recently, the issue has also been taken up by summarization research (Barzilay, 2003; Ji and Pulman, 2006; Madnani et al., 2007). In the spirit of the latter, this chapter investigates the following three questions:

1. Which factors are most important for determining coherent sentence orderings?

2. Does the topic of the text influence the factors that are important for sentence ordering?

3. How much performance is gained when using deeper processing/knowledge resources?

This chapter is structured as follows. Section 6.2 presents previous research relating to the optimal ordering of sentences, both as an individual task (i.e., as a simplification of the natural language generation component) and in the context of automatic summarization. Section 6.3 presents my model for the sentence ordering problem as well as an exploration of the various attributes that I consider important to determine the quality of the ordering. Section 6.4 describes experiments for selecting and evaluating the model and compares the present results with those obtained via state-of-the-art approaches. Section 6.5 concludes the chapter by discussing the results and outlining potential future work.

6.2 Related Work

6.2.1 A Simple Classification

The ordering of information in summaries created by means of automatic summarization systems has been approached from a multitude of directions. One perspective, knowledge-rich approaches, relies on manually creating representations of sentence orderings using domain communication knowledge (Rambow, 1990; Kittredge et al., 1991). Formalisms using these types of knowledge are scripts (DeJong, 1982), schemata (McKeown, 1985), and domain-specific schemata (Rambow, 1990). The relevant approaches are based on the idea that people have preconceived notions regarding the means with which communicative goals can be achieved as well as the integration of these means to form a text; in other words, text reflects one or more principles of organization. One such notion, in the context of a narrative, is that the narrative should begin with a description of the setting, which usually includes an overview of the characters, scene, and time frame. In the case of McKeown (1985), the strategic component – i.e., the component that determines the content and structure of the discourse (generated using natural language generation on database knowledge based on a query) – uses four manually determined patterns called schemata: identification, constituency, attributive, and contrastive. These schemata are described using a grammar that incorporates both optionality and repetition.

At the other end of the spectrum are knowledge-lean approaches, which do not utilize manually inferred knowledge about the target domain. Instead, they attempt to

discover automatically the underlying discourse structure. The approaches discussed in the remainder of the section are mainly based in this paradigm.

While the foregoing classification distinguishes between manually or automatically exploiting knowledge about a particular domain or situation, past research has studied a plethora of other issues pertaining to sentence ordering. To place the methodology proposed in this chapter into its broader context, this section provides a brief overview of the most prominent approaches put forward in the past. An in-depth discussion of the aspects pertinent to the present work is provided in Section 6.3.

6.2.2 Sentence Ordering Based on Chronology

In general, the order of presentation of information in any kind of text is determined by a vast number of factors. Examples include author preference, target audience, and the informational focus of the presentation. Opportunely, a number of relevant factors tend to follow common patterns and are therefore relatively tractable. Examples of such factors include the timeline of the source documents, sequence of topics, and lexical and syntactic links between different sentences. The following explores a number of different avenues that can be pursued in order to determine the best ordering of a set of sentences involving one or more of these attributes.

One approach to sentence ordering in the context of automatic summarization is the ordering of the sentences in the summary according to the publication date of the documents from which the summary sentences are selected (McKeown et al., 1999; Barzilay et al., 2002). The basic idea is that the summary presents the information in the same order in which they occur in the source documents. A central aspect in this regard is that one has to take into consideration that the same unit of information may occur in multiple documents. Thus, it is important to have some rudimentary measure of similarity for units of information.

A related approach is the incorporation of textual cues regarding the temporal ordering of sentences (Bollegala et al., 2006). Approaches of this kind are also based on the chronological ordering of the information. However, they refine the notion such that actual textual cues in the sentences are exploited, rather than the publication date being used as an approximation. If one sentence talks about some event in 1995 while another refers to an event in 1989, the ordering according to these approaches would first present the sentence containing 1989 and then the sentence containing 1995. Although intuitive, one of the main problems with this approach is that not all sentences

contain textual cues, which makes it difficult to use as an exclusive tool.

Another class of approaches is based on one of the most prominent techniques in single-document summarization – the use of the ordering of information in the source documents. When translating this approach to a multi-document summarization system, however, one encounters a number of problems. First, one needs to have some sort of measure for determining whether two sentences are sufficiently similar to be considered in the ordering. Second, one needs to be able to determine which ordering is optimal. The key difficulty in the context of the latter is that there frequently exist conflicting orders of the same two units of information in different source documents. The upshot is that determining the actual ordering is a complex problem.

Barzilay et al. (2002) use an approximation algorithm and find that the ordering obtained using this method is good if the underlying source documents exhibit high agreement with each other with regard to the ordering of the information. The results are poor if there is no clear order indicated by the source documents. In these cases, the algorithm frequently determines an order of the sentences that did not occur in the source documents, which according to their results is problematic since these orderings tend to achieve the lowest scores. Owing to problems with the pure chronological ordering of sentences, they (Barzilay et al., 2002) subsequently extend their chronological approach in such a way that topically related sentences are grouped together; a fact they observed in a human study. As part of this approach, they segment the texts based on word distribution and co-reference analysis and determine whether sentences of the same topics tend to co-occur in the same text segments within the individual documents. If they co-occur often, this is taken to indicate that the sentences should be grouped together. In the next step, they assign each block of topics a time stamp (the earliest time stamp of the topics the block contains). To arrive at an ordering, they then apply their chronological ordering approach to the time-stamped blocks. The consequence of this more topic-oriented approach is a notably improved performance.

6.2.3 Sentence Ordering Based on Non-Temporal Cues

The remainder of the approaches described in this section completely disregard the order of the information in the source sentences or temporal cues, either by design choice or because the relevant approaches do not address the problem of sentence ordering in the summarization context. Barzilay and Lee (2004) base their approach on the notion of content, which roughly corresponds to the notion of topic explored

above. In essence, they want to capture the characteristics of the orderings of sentence topics in specific domains using an automatically induced schema-like structure. To be more precise, they aim to discover automatically the topics of sentences and the ordering of the topics in the whole domain-specific source documents. Their approach involves the following sequence of processing steps:

- All texts in a given domain are generated by *one* content model, a Hidden Markov Model (HMM; Rabiner, 1989), where each state corresponds to a distinct topic and generates sentences based on a state-specific language model. State transition probabilities capture constraints on topic shifts.

- The initial topic induction is achieved via complete-link clustering with k clusters in conjunction with cosine similarity using bi-grams as features. Under the assumption that not all sentences conform to particular clusters, but may discuss irrelevant or new content, all clusters containing less than T sentences are merged into one cluster, named the "et cetera" cluster. As a result, there are m content clusters.

- An HMM containing m states is constructed, where each state corresponds to one of the clusters from the initial topic induction. For each of the states, except the "et cetera" state, the language model is generated using the cluster members. The last state's language model is computed to be complimentary to those of the other states. The state transition probabilities are computed using the number of transitions from one topic to the next.

- The EM-like Viterbi (Iyer and Ostendorf, 1996) approach is used to re-estimate cluster membership based on the most likely language model to have generated the sentence. The re-estimation is repeated until cluster membership stabilizes.

The result of this procedure is a score for each document (ordering of sentences) that specifies the likelihood of the sentence ordering according to their HMM, i.e., the likelihood of the transition between sentence topics and the likelihood of generating the sentence for that topic. They evaluate their approach on a number of different domains and find enormous differences between them. One of their evaluations compares the score achieved by the original ordering to the scores of alternative re-orderings of the original sentences. They look at the relative position of the original ordering among the alternative orderings and find that the variability is high. For example, in the finance

domain, the original ordering is among the top 5 orderings in 100% of the cases, while this is only true in 47% of the cases in the drug domain.

An alternative class of approaches is based on theories of coherence, i.e., theories using syntactical structure, logical tense structure, presuppositions, and implications connected to general world knowledge. One such theory is Centering Theory (Grosz et al., 1995). Barzilay and Lapata (2008) use the basic assumptions of centering theory – that some entities in text are more important than others and that their syntactic occurrence and position influences the choice of referring expression – to create entity grids that represent the existence and syntactic position of entities in sentences, which are then translated into features for their machine learning algorithm.

In more detail, they parse the text in order to obtain syntactic information for the noun phrases; in particular, whether a noun phrase is the subject, object, or neither of a sentence. They then arrange the noun phrases and their occurrences in different sentences in an entity grid. Table 6.1 provides an example. Table 6.1a contains the sample text. The [] annotations followed by subscripts denote the noun phrases and their syntactic role. The arrangement of these phrases into entity grids is illustrated in Table 6.1b. The six rows in the table correspond to the sentences in the sample text. Each column represents the sequence of words in the different sentences. At this point, they utilize local entity transitions (inspired by the local transitions in Centering Theory), which are represented as a sequence $\{s,o,x,-\}^n$ that represents entity occurrences and their syntactic role in n adjacent sentences. For each local entity transition, they compute the probability of the transition in the document. They then represent each document as a feature vector containing the local entity transitions. As a final step, they utilize machine learning – support vector machines (SVMs) in particular – in order to learn a model that captures the importance of the transition sequences.

6.2.4 Sentence Ordering Based Purely on Machine Learning

While the approaches to sentence ordering thus far had a relatively refined view of the properties that impact on the ordering of sentences or the coherence of text, the approaches presented in this section do not adopt such views. Instead, they use a high-dimensional feature space and use either machine-learning or dimension-reduction approaches to arrive at an ordering of the sentences.

Even though similar to Barzilay and Lapata (2008), Lapata (2003), for example, views the ordering task in a slightly different manner. Both use machine learning, yet

1	[The Justice Department]$_{\mathbf{S}}$ is conducting an [anti-trust trial]$_{\mathbf{O}}$ against [Microsoft Corp.]$_{\mathbf{X}}$ with [evidence]$_{\mathbf{X}}$ that [the company]$_{\mathbf{S}}$ is increasingly attempting to crush [competitors]$_{\mathbf{O}}$.
2	[Microsoft]$_{\mathbf{O}}$ is accused of trying to forcefully buy into [markets]$_{\mathbf{X}}$ where [its own products]$_{\mathbf{S}}$ are not competitive enough to unseat [established brands]$_{\mathbf{O}}$.
3	[The case]$_{\mathbf{S}}$ revolves around [evidence]$_{\mathbf{O}}$ of [Microsoft]$_{\mathbf{S}}$ aggressively pressuring [Netscape]$_{\mathbf{O}}$ into merging [browser software]$_{\mathbf{O}}$.
4	[Microsoft]$_{\mathbf{S}}$ claims [its tactics]$_{\mathbf{S}}$ are commonplace and good economically.
5	[The government]$_{\mathbf{S}}$ may file [a civil suit]$_{\mathbf{O}}$ ruling that [conspiracy]$_{\mathbf{S}}$ to curb [competition]$_{\mathbf{O}}$ through [collusion]$_{\mathbf{X}}$ is [a violation of the Sherman Act]$_{\mathbf{O}}$.
6	[Microsoft]$_{\mathbf{S}}$ continues to show [increased earnings]$_{\mathbf{O}}$ despite [the trial]$_{\mathbf{X}}$.

(a) A representative text fragment augmented with syntactic annotations for grid computation. (Reproduced from Barzilay and Lapata (2008), Table 2.)

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	**S**	**O**	**S**	**X**	**O**	–	–	–	–	–	–	–	–	–	–	1
2	–	–	**O**	–	–	**X**	**S**	**O**	–	–	–	–	–	–	–	2
3	–	–	**S**	**O**	–	–	–	–	**S**	**O**	**O**	–	–	–	–	3
4	–	–	**S**	–	–	–	–	–	–	–	–	**S**	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	**S**	**O**	–	5
6	–	**X**	**S**	–	–	–	–	–	–	–	–	–	–	–	**O**	6

(b) A fragment of the entity grid for the sample text in Table 6.1a. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical role: subjects (s), objects (o), neither (x). (Reproduced from Barzilay and Lapata (2008), Table 1.)

Table 6.1: A sample text annotated for entities and the associated entity grid.

Lapata (2003) does not directly assess the quality of the ordering and instead learns to predict the next sentence given the current sentence, which represents a coarse attempt at capturing Marcu (1997)'s local coherence constraints and Barzilay et al. (2002)'s observations about topical relatedness. The features she uses are derived from three categories – verbs, nouns, and dependencies – all of which are lexicalized. With regard to verbs, the features capture the verbs' lemmatized forms as well as versions that retain tense information. For nouns, she uses features for all nouns (ignoring pronouns, however) and utilizes these features as approximations to entity-based local coherence. The last set of features, dependencies, is obtained using Minipar (Lin, 1998). Different versions of her algorithm use different types of dependencies (those including nouns, verbs, adjectives, adverbs, and prepositions). Using this set of features, Lapata (2003), to some degree, learns a kind of precedence between the words and features in the sentences, which in turn also represent topics. However, a comparison of the results of Lapata (2003) and Barzilay and Lee (2004) suggests that explicit modeling of topics is preferable to implicit modeling, as suggested by the better results obtained by Barzilay and Lee (2004).

While Lapata (2003)'s approach uses linguistic knowledge, Foltz et al. (1998) do not use linguistic knowledge beyond what constitutes a word. Instead, they use latent semantic analysis (LSA; Deerwester et al., 1990), which uses a term-document matrix to represent the occurrence of terms in documents. Then a low-rank approximation of this matrix is computed, typically using singular value decomposition to obtain a set of orthogonal eigenvectors. Using fewer eigenvectors than the dimensionality of the original matrix and replacing the other eigenvectors by $\vec{0}$, an approximation of the original matrix in a lower dimension is obtained. By way of this approximation, semantic relations between terms can be obtained, i.e., term dimensions are combined and the eigenvectors are linear combinations of the original term dimensions. Using this matrix of eigenvectors, Foltz et al. (1998) map individual sentences (represented as term vectors) into the lower-rank space. They then compare two sentences in the lower-rank space using cosine similarity and apply a measure of similarity (e.g., cosine similarity) of two sentences to determine the quality of their sentence ordering. This approach achieves good results using an extremely knowledge-lean approach (judged as per the comparison of the results of different approaches in Barzilay and Lapata (2008)).

6.2.5 Discussion: Sentence Ordering

This section illustrated a number of general approaches to sentence ordering that exploit a wide variety of sources of information, ranging from manual to automatic approaches, from syntactically to topically motivated approaches, as well as from distributional to temporal-order approaches. A broad comparison of the different techniques indicates that there are quite a few differences relating to the attributes of text considered important for the creation of coherent text or, more specifically, the ordering of the sentences that results in the most readable/understandable text. Barzilay and Lapata (2008), for instance, rely merely on the presence/absence of entities and syntactic cues in the sequence of sentences without any regard for the content of the sentences, while Barzilay and Lee (2004) are only interested in the content and the content transitions. Others, in turn, utilize features inherent in the source documents from which the sentences are selected to construct an ordering. These considerations show that, in principle, a multitude of approaches are available for the objectives at hand.

I intend to investigate an approach that utilizes entity transitions; at the same time, I consider the relative impact of the similarity of sentences, which provides an alternative measure of the sequence of sentences. In theory, this combination of attributes enables the approach to capture both syntactic properties of the sentences to be ordered and general topic shifts. The idea is that exploiting both ideas should provide for a better generalization of the model as it captures a larger number of (different) patterns of document structure than do existing approaches.

Although also based on machine learning, the present work takes a completely different approach than Lapata (2003) insofar as I do not attempt to learn the ordering preferences between pairs of sentences. Rather, I compute scores for documents using individual features, which are then combined to a total score using the preference rankings in the training set. The advantage of this approach to sentence ordering is that it allows one straightforwardly to discern the individual value of the features investigated in this chapter. Likewise, in comparison to Barzilay and Lapata (2008), my approach uses shallower features. Moreover, from a computational perspective, my features are based on pairs of sentences and thus do not require re-computation of all features, but only the selection of correct values for the given feature vectors. Barzilay and Lapata (2008)'s feature values, in turn, involve more than two sentences for entity sequences.

Note that this section did not spend much time on the features utilized by the various machine learning approaches. Where appropriate, the features and feature repre-

sentation are explored in more detail in Section 6.3.2. The following section introduces my knowledge-lean approach to determining a good ordering for a set of sentences by describing the machine learning approach, the data representation, and the features to be explored.

6.3 The Model and Feature Representation

6.3.1 The Model

I view sentence ordering as a machine learning problem that I approach using the feature representation discussed in detail in Section 6.3.2. In general, looking at the sentence ordering problem, it is very difficult to assign absolute quality scores to a specific ordering of sentences because there is no absolute best ordering, though certain orderings are often preferable (say, depending on the information requirement(s)). Deciding which out of a pair of orderings is preferable is an easier task since it only requires a decision between two given orderings as opposed to the exploration of all possible orderings and picking one for the full sentence ordering problem. I therefore view sentence ordering as a ranking task that needs to assess whether one ordering of the same set of sentences is more coherent than another. Given this view, my model has the following overall structure (which parallels the approach by Barzilay and Lapata (2008)).

Data. The data for the machine-learning approach consists of alternative orderings (x_{ij}, x_{ik}) of the sentences of the same document (d_i). In the training data, the partial preference ranking of the alternative orderings is known. To be more precise, the ordering between the original ordering and each of the alternative orderings is known; there is, however, no knowledge regarding the ranking between two alternative orderings. As a result, training consists of determining a parameter vector $\vec{w}$ that minimizes the number of violations of pair-wise rankings in the training set.

Machine Learning. The training problem can be solved using a number of different machine learning approaches. Among them is SVM constrained optimization (Joachims, 2002), a technique used successfully in a variety of natural language processing tasks, including search engine optimization and parsing (Joachims, 2002; Toutanova et al., 2004). Paralleling Barzilay and Lapata (2008), this is the technique I

```
1 qid:1 1:0.5 2:4 3:5
0 qid:1 1:1.5 2:2 3:1.5
0 qid:1 1:1 2:1 3:3
0 qid:1 1:0 2:0 3:1
1 qid:2 1:0.5 2:4 3:1
0 qid:2 1:0 2:2 3:3
0 qid:2 1:1.5 2:3 3:1
0 qid:2 1:1.5 2:0 3:4
```

Figure 6.1: An Example of the learning file of a ranking SVM.

use – the SVM^{light} implementation[1] of the ranking SVM, to be specific.[2]

In practical terms, an SVM file would look like the (excerpt of a) learning file depicted in Figure 6.1. Each line represents an individual ordering of a set of sentences. In this particular example, there are two documents, identified by qid:1 and qid:2, with four orderings of the sentences each (denoted by the four lines with the same qid). The correct ordering is assigned a target value of 1 (the first number in a given line), while all alternative orderings have a target value of 0. The remaining number pairs identify feature-value pairs. In ranking mode, SVM^{light} uses the target values to determine pair-wise preference constraints subject to the qids, based upon which it then optimizes Kendall's tau (for a detailed derivation of the ranking SVM algorithm refer to Joachims (2002)).

Figure 6.1 demonstrated the *document*-level data available to the machine-learning algorithm and how it is used in the computation of the ranking model. Yet, the features described in Section 6.3.2 relate to pairs of *sentences*. As such it is necessary to aggregate the information on a document-level for each of the individual features. To this end, the minimum, average and maximum feature scores between sentences were investigated. Minimum and average scores showed very little differences during the course of the explorations. Maximum scores, on the other hand, generally performed quite poorly, although this is hardly unexpected since the maximum does not take into account any other sentence-pair scores. For this reason, in the following explorations, only the average method is used to derive document-level scores from sentence-level scores for the features. For example, consider a document with four sentences. Assum-

[1] `http://www.cs.cornell.edu/People/tj/svm_light/`

[2] For a detailed derivation of the mathematical basis for this approach, refer to Boser et al. (1992) for classification SVMs and Joachims (2002) for ranking SVMs.

ing that there are two entity matches between sentences 1 and 2, one match between sentences 2 and 3, and two between sentences 3 and 4, the document-level score for this feature would thus be 5/3 – five entity matches divided by three sentence-pairs.

Evaluation. The quality of a particular ordering is obtained by averaging the scores obtained for each pair of adjacent sentences. Given the scores of two orderings of the same set of sentences, the higher scoring document is the document that has the better ordering.

Because of the use of machine-learning, the remaining task is the selection of appropriate features for determining the preference between alternative orderings. The following section is dedicated to this task.

6.3.2 Features

Section 6.2 described a number of different approaches not based on the explicit modeling of the sentence order, which instead extract relevant information from a training set of documents. The features used for modeling by the relevant systems included language models (n-gram models) (Barzilay and Lee, 2004), lexicalized features (Lapata, 2003), chronological features (Barzilay et al., 2002), word distribution (Barzilay et al., 2002), as well as syntactic knowledge of entity distributions (Barzilay and Lapata, 2008).[3]

The features I use involve two different knowledge resources: WordNet (Fellbaum, 1998) and VerbOcean (Chklovski and Pantel, 2004). Recall that WordNet is a large lexical database of English that organizes nouns, verbs, adjectives, and adverbs into a set of cognitive synonyms that are interlinked by means of lexical and conceptual-semantic relations. VerbOcean, on the other hand, is broad-coverage repository of semantic relations between verbs. The relations – similarity, strength, antonymy, enablement, and temporal "happens-before" – are automatically mined from data on the web. I am most interested in the "happens-before" relation, which indicates whether "two verbs refer to two temporally disjoint intervals or instances" (Chklovski and Pantel, 2004), as it provides an alternative means for obtaining chronological information if no explicit textual cues in the text indicate different timeframes. Examples of rela-

[3]Most of the sentence ordering approaches presented in this chapter deliberately used relatively shallow features, because I initially intended to use the quality of the given sentence ordering as a component for the sentence selection process as alluded to in the introduction to this chapter. However, due to time constraints, the experiment now uses sentence ordering as a post-processing step to create readable summaries.

tions in the VerbOcean repository are marry – divorce, detain – prosecute, and enroll – graduate. One clearly has to be married in order to get divorce from one's spouse. Likewise, one has to be enrolled at a school or university before one can graduate from the institution.

The specific sentence ordering technique(s) developed in this chapter can be grouped into four categories:

- **Unit size.**
  In terms of features, Section 6.2 described approaches based on whole sentences (Barzilay and Lee, 2004; Barzilay et al., 2002) as well as approaches that only consider noun phrases (Barzilay and Lapata, 2008). To determine which unit works best for determining the order of a set of sentences, I investigate which unit is most helpful with respect to discovering the sequence of a set of sentences. In particular, I explore the following units: sentence, noun groups, verb groups, heads of the noun and verb groups, and the combination of the heads of the noun and verb groups. The different chunks are determined using LT-TTT2 (Grover and Tobin, 2006). The investigation of the appropriate unit size is partially inspired by discourse entity-based accounts of local coherence (e.g., Kuno, 1972; Halliday and Hasan, 1976; Grosz et al., 1995). Yet, in contrast to Barzilay and Lapata (2008)'s syntactic considerations, I only consider whether or not the entities occur in adjacent sentences. I also investigate whether information is gained when using whole noun groups as opposed to the head words of the noun groups (only).

- **Similarity measure for units.**
  Even if appropriate unit size for the comparison of sentences is determined, it is still necessary to determine whether two units in two different sentences are similar. I use the following information to this end: surface form matching, lemma matching, and matching based on various WordNet relations – synonym, hypernym, hyponym, and antonym in particular. One would expect that use of the various different relations would improve the matching between sentences. At the same time, however, they might find excessive matches between unrelated sentences because of ambiguity.

- **Temporal relations.**
  As indicated above, the use of information regarding temporal cues has provided good results in the sentence ordering for summarization systems (if combined

with other measures). The techniques based on VerbOcean exploited in this category strive to facilitate the use of a new knowledge resource to determine chronological orderings more accurately.

- **Size of local context.**
  The techniques discussed thus far only concerned relations between directly adjacent sentences. However, within a certain topic (expressed across a number of sentences), one would expect higher similarity between the sentences on the same topic than between other sentences. To explore the larger relative context of sentences (which Barzilay and Lapata (2008) exploit via their entity sequences), I explore the similarity of the sentences within windows of n sentences.

In sum, the technique proposed in this chapter is based on SVM machine learning and involves a feature representation involving chunk information, lexical semantics, chronological information, and similarity between non-adjacent sentences. It is expected to yield a good performance while, at the same time, providing good generalization capabilities because of its shallow features. The remainder of this chapter strives to engineer the optimal combination of features to the end of achieving the best possible text structure.

6.4 Experiments

The approach described in the previous section is evaluated using the synthetic datasets developed and used by Barzilay and Lapata (2008). The rationale for using their datasets is comparability to other state-of-the-art systems. Furthermore, I follow Barzilay and Lapata (2008)'s experimental procedure insofar as evaluation on the initial datasets and cross-training between their two datasets is concerned. A third dataset, based on DUC2005 data, is introduced in order to evaluate the impact of the topic of the texts on general performance. For transparency, Section 6.4.1 discusses the evaluation framework in more detail, including the different datasets, whereafter Sections 6.4.2, 6.4.3, and 6.4.4 present the specific results derived from my model of sentence ordering.

6.4.1 Evaluation and Datasets

An interesting and important general realization regarding the evaluation of sentence orderings is that there is often no unique best ordering of the sentences extracted by

automatic summarization systems (Barzilay et al., 2002). Madnani et al. (2007), moreover, show that the set of coherent orderings in newswire text is larger than in other types of text. Hence, given a set of sentences extracted from any text, it is impossible to determine a unique optimal ordering, particularly so in the case of newswire text.

This suggests that the automatic evaluation of sentence ordering algorithms is not necessarily straightforward in the sense that there is no 'most' correct ordering, but there may be a number of coherent orderings – especially in the case of automatically selected sentences in summarization systems. To avoid this problem, the three datasets used for the automatic evaluation in this chapter are based on human-generated texts, for which the intended ordering is often more apparent (on the assumption that human text is inherently coherent).

The first two datasets are the earthquake and accident datasets used by Barzilay and Lapata (2008), which are not related to summarization in any way. Each dataset consists of pairs of the original text ordering and a random permutation of the sentences; for each text, the dataset contains 20 random permutations. In total there are about 2,000 pairs in the training and test sets of each dataset. Figure 6.2 provides a sample of a typical document (original ordering of the sentences) in the accident dataset. The documents typically start with a header, followed by a one-sentence summary, the sequence of events that led up to the accident, and a description of the damage, respectively. Figure 6.3 provides a sample document (original ordering of the sentences) from the earthquake dataset. These documents typically start with a headline, followed by a summary sentence, facts about the earthquake, and the potential damage caused, respectively.

The third dataset is similar to the first two in that it contains original texts and random permutations. In contrast to the other two sources, however, this dataset is based on the human summaries from DUC2005 (Dang, 2005). In particular, it contains 300 human summaries on 50 document sets (4 or 7 summaries per document set), resulting in a total of 6,300 documents split into training and test sets (i.e., a total of 6,000 pairs of documents containing the original ordering and one of 20 permutations ($20 \cdot 300 = 6,000$)). The dataset furthermore differs from Barzilay and Lapata (2008)'s datasets in that the content of each text is not based on one individual event (an earthquake or accident along with the number of victims and damage), but on more complex topics followed over a period of time (e.g., the development of the relations between Britain and Argentina with regard to the Falklands conflict). Figure 6.4 provides a sample document from this dataset. The particular document deals

This is preliminary information, subject to change, and may contain errors.
Any errors in this report will be corrected when the final report has been completed.
On January 13, 1994, about 1230 hours Greenwich Mean Time, a Beech BE-90, N46WA, registered to Charles Kuykendall, Wilmer, Texas, was destroyed during a ditching in international waters about 50 nautical miles south of Martigues, France.
The ditching was precipitated by an in-flight fire during cruise flight.
The German national commercial pilot, the sole occupant, received minor injuries.
Visual meteorological conditions prevailed and a flight plan was not filed.
The ferry flight departed from Straubing-Wallmuhl, Germany.
According to the French Bureau Enquets Accidents and the FAA, the airplane was being ferried from Germany to Dallas, Texas, with a planned stop in Portugal.
According to the pilot, an electrical fire started and produced smoke in the cockpit during the flight from Straubling-Wallmuhl, Germany, to the Azores, Portugal.
The pilot elected to ditch the airplane into the Mediterranean Sea.
After the ditching, the pilot utilized survival equipment and was later rescued by a search and rescue helicopter.
The airplane sank.
According to Hungarian records, the airplane was substantially damaged in a previous accident on March 9, 1991.
The airplane remained in Hungary until it was purchased by the current registrant.

Figure 6.2: A sample document from Barzilay and Lapata (2008)'s accident dataset.

BC-Greece-Eathquake|Minor Quake Shakes Northern Greece
SALONICA, Greece (AP) A minor earthquake shook this northern port city on Monday, but caused no damage, injuries or panic.
The Salonica Seismological Institute said the quake had a preliminary magnitude of 4.7 and that its epicenter was located 18 kilometers (11 miles) northwest of this city of one million people.
It occurred at 3:16 p.m. (1216 GMT).
Police reported no damage, injuries or panic and said the quake was felt throughout the region.
They said city schools were evacuated as a precaution.
A quake of moment magnitude 2.5 to 3 is the smallest generally felt by people while a temblor of magnitude 4 often causes slight damage.
Magnitude 5 can produce moderate damage.
Magnitude 6 causes severe damage under a populated area.
Magnitude 7 indicates a major earthquake capable of widespread, heavy damage.
Magnitude 8 is a "great" earthquake capable of tremendous damage.

Figure 6.3: A sample document from Barzilay and Lapata (2008)'s earthquake dataset.

with the British-Argentinian relations regarding the Falkland Islands. It shows that the documents from this dataset are significantly more complicated and contain more information than the documents in the other two datasets. Since the different document sets cover completely different topics, my learning algorithm has to sidestep the issue of topic-dependence (a typical problem of machine learning approaches) in order to perform well on this dataset. The third dataset will thus mainly be used to evaluate topic-independent properties of the approach presented in this chapter. Table 6.2 summarizes the information for the datasets.

Dataset	**Training**	**Test**
Earthquakes	1,896	2,056
Accidents	2,095	2,087
DUC2005	up to 3,300	2,700

Table 6.2: The datasets. The table provides an overview of the number of pair-wise rankings in the training and test sets within each of the three datasets used for the experiments.

Britain resumed trade with Argentina in 1985. In 1989 Argentina welcomed British imports.
In 1990 diplomatic ties resumed.
Britain lifted military protection zones around the islands. Each side gives advance notice of military exercises.
Argentines visit war cemeteries in the Falklands. Britain refutes Argentina's claims of sovereignty over the Falklands, disputed since 1833.
Argentina's main objective remains recovering sovereignty of the Falklands, thinking economic links would reduce the Falkland's importance to Britain.
Argentina "respects" the islands' "history", but won't recognize its local government.
Britain's arms embargo continues, except for Argentine units participating with Gulf War Coalition forces.
Argentine military officers train in the UK.
Britain won't grant Argentine President Menem's request to visit the UK.
Ministers have been exchanged.
After 1990, rapid UK trade growth helped Argentina's "miracle" economic recovery.
Argentina wants UK regulatory expertise in privatization and private sector investment in Argentina's gas and nuclear industries.
British Gas bought the largest Argentine gas company and exploits off-shore gas with Argentine companies.
A 1993 fisheries conservation agreement changed Argentine policies which threatened depletion of fragile fish populations.
Britain angered Argentina by extending territorial waters to 200-miles around South Georgia and South Sandwich islands, which Argentina claims, and when it extended fisheries control within the Falkland's 200-mile limit.
Argentina briefly banned UK cattle imports and opened investigations into allegations of Falkland War atrocities by British soldiers.
In 1993 British business insurance covered UK companies in Argentina.
In 1994 Britain indicated readiness for Argentine companies to participate in Falkland's off-shore oil development.

Figure 6.4: A sample document from the third dataset – a human summary from document set D324 (DUC2005).

6.4.2 Experiment 1: Generic Sentence Ordering

The first experiment has five main objectives. First, it investigates the impact of different granularities of shallow syntactic units. That is, does it make a difference whether whole sentences, only noun groups, or only heads of noun groups are considered in the sentence ordering process? Second, the effect of WordNet synonyms, hypernyms, hyponyms, and antonyms is assessed. Third, it explores the usefulness of temporal relations provided by VerbOcean. Fourth, while local models of coherence typically only consider direct sentence-to-sentence aspects of coherence, I also investigate the benefit of allowing for slightly longer-range relations, e.g., to the sentence preceding the previous sentence. Lastly, this experiment provides a reference point to compare my local, shallow, non-lexicalized approach to deeper syntactically motivated (Barzilay and Lapata, 2008), global (Barzilay and Lee, 2004), and lexicalized (Foltz et al., 1998) models of sentence ordering.

Note that within the latter set of approaches, the shallow syntactic unit of analysis varies considerably. While Foltz et al. (1998) consider individual words for latent semantic analysis, Barzilay and Lee (2004) use a bi-gram language model; both use the whole sentence as a unit of comparison. Barzilay and Lapata (2008), in contrast, use noun groups and co-reference information between noun groups as the unit for determining whether or not two sentences are coherent. Since these approaches (per se) are vastly different, I seek to determine the superiority of a specific unit of comparison. Whereas Barzilay and Lapata (2008) implicitly consider some of the aspects regarding the units of comparison in their co-reference resolution algorithm, I strive to capture *explicitly* the importance of particular units on the overall coherence of a document.

Exploration 1. To these ends, I use five (different) units as the basis of my coherence analysis: sentence, noun group, head of the noun group, verb group, and head of the verb group. These units are obtained by processing the documents using the LT-TTT2 tools (Grover and Tobin, 2006); the lemmatizer used by LT-TTT2 is morpha (Minnen et al., 2000). The results of these initial considerations are shown in Table 6.3.

One of the most obvious results of this exploration is that sentence-level analysis does not perform well. The apparent explanation for this finding is the impact of stop-words on the similarity between sentences; however, an alternative implementation using tf.idf-weighted words did not improve the results.[4] It would thus appear that stop-word removal is the only way to boost performance for this syntactic category. I

[4]The relevant results are not reported since no more information was gained beyond the fact that overall performance is poorer than when using the method reported.

Syntactic Unit	**Processing**	**Accuracy**	
		Accidents	***Earthquakes***
sentence	surface form	52.27	14.21
	lemma	52.27	12.04
heads sentence	surface form	77.35	60.30
	lemma	73.18	61.67
noun group	surface form	80.14	59.84
	lemma	81.58	59.54
head NG	surface form	80.49	59.75
	lemma	**81.65**	59.12
verb group	surface form	71.57	68.14
	lemma	53.40	68.01
head VG	surface form	71.15	**68.39**
	lemma	53.76	67.85

Table 6.3: Results of Experiment 1 (Part 1). Performance with respect to the syntactic unit of processing of the training datasets. "Accuracy" is measured as the fraction of correctly ranked pairs divided by the total number of pairs for which a ranking was obtained. Note that "heads sentence" denotes the union of the heads of the noun and verb groups as opposed to the head of the main verb group.

did not investigate this avenue further as the performance of the "heads sentence" unit did not perform well either. Since this category only contains the heads of the noun and verb groups, the poor performance can only be explained by the mixing of verb and noun information.

An interesting result is the difference in performance between the surface and lemmatized versions for verb syntactic units. The higher performance in the case of surface forms can most readily be attributed to the tense and person information encoded in the surface forms. In the earthquake domain, in particular, this information appears to be more important than the features related to discourse entities. In fact, the surface forms encode similar information as Lapata (2003)'s verb features that retain temporal information. For example, in the earthquake domain, the use of tense can indicate shifts from the description of the earthquake that occurred in the *past* to the *current* efforts to ameliorate the conditions or damage that resulted from the earthquake. Even if that shift is not available, sentences with the same tense are mostly grouped together (*cf.* Figure 6.3, where the earthquake and damage are in past tense, while the general description of earthquake magnitude is in present tense).

Manual inspection of the two datasets reveals that the major overall differences in performance can be ascribed to two main factors: (1) The accidents dataset is composed of official accident reports for aircrafts, which follow quite clearly the events of the accident, and (2) the earthquakes dataset comprises plenty of sentences containing damage reports, which have no relation to one another apart from the fact that they describe some sort of damage; these sentences can often be ordered arbitrarily.

Exploration 2. Moving from purely syntactic considerations to semantic ones, WordNet, in principle, provides a useful resource for determining semantic similarities and relations between lexical representations. In practice, however, it can frequently not be successfully exploited in many application scenarios because of ambiguity in the lexical representations, which result in a rapidly exploding, inaccurate representation of meaning. Since the problem in sentence ordering research is the identification of *any* antecedents that are in some way related to the current sentence, the increase in possible meanings may, in fact, be beneficial as it captures information that is similar to the information obtained when using dimensionality reduction in latent semantic analysis, though based on a more accurate source.

One of the fundamental reasons for the usefulness of semantic resources is that lexical repetitions in texts are considered bad style, resulting in the use of synonyms or other semantically similar expressions. With this background, the aim of this part of

Syntactic Unit	Processing	Accuracy	
		Accidents	*Earthquakes*
noung group	synonyms	82.06	59.74
	hypernyms	75.94	61.54
	hyponyms	80.56	59.94
	antonyms	74.03	48.41
	combined	72.22	56.86
head NG	synonyms	**82.37**	59.40
	hypernyms	76.98	61.02
	hyponyms	81.59	59.14
	antonyms	74.20	48.07
	combined	70.84	56.51
Sentence	synonyms	65.04	40.80
	hypernyms	60.35	49.70
	hyponyms	64.76	40.69
	antonyms	56.72	66.47
	combined	53.51	50.44
heads Sentence	synonyms	70.96	61.04
	hypernyms	59.51	58.62
	hyponyms	65.65	56.97
	antonyms	56.63	60.83
	combined	55.45	57.0
verb group	synonyms	54.29	**70.84**
	hypernyms	52.89	61.16
	hyponyms	55.49	68.28
	antonyms	47.47	63.65
	combined	50.15	66.39
head VG	synonyms	54.19	70.80
	hypernyms	53.36	60.54
	hyponyms	55.27	68.32
	antonyms	47.45	63.91
	combined	49.73	66.77

Table 6.4: Results of Experiment 1 (Part 2). The impact of WordNet on coherence accuracy in the training datasets.

the experiment is to determine (1) whether WordNet helps in creating more coherent documents, and (2) which WordNet relations are (most) helpful/harmful with respect to the performance in the sentence ordering task, e.g., synonyms, hypernyms, hyponyms, antonyms.

The results, presented in Table 6.4, provide a varied picture. In the accidents dataset, synonyms and hyponyms generally, and hypernyms occasionally, provide good results. Antonyms and the straightforward combination of all four relations, however, do not seem to increase performance. The results for the earthquakes dataset are similar, though there is an increase in performance of noun group-based features (as compared to the results without WordNet reported in Table 6.3). Nevertheless, compared to the accident dataset, the performance of noun group features is still poor.

The unexpectedly poor performance of the combination of all four relations either indicates an overgeneralization in the semantics, or is due to the equal weight assigned to each relation. For this reason, the WordNet features will utilize the four relations individually and allow the machine-learning algorithm to obtain optimal weights for each. Overall, sentence ordering performance slightly increases when using WordNet compared to the results in the initial experiment that only considered surface forms and lemmas. Yet, for the units for which lemma features performed worse than surface forms, the use of WordNet only provides results that are comparable to the lemma based comparison, i.e., they do not achieve results that are similar to those achieved using surface forms. For the accident dataset, the best result improved by about 0.5% compared to the best result in the preceding experiment, while the increase for the earthquake dataset is approximately 2% in the best case.

Exploration 3. The third category of features investigated as part of this experiment is the temporal ordering of sentences. I use VerbOcean to obtain the temporal precedence between two events (denoted by the main verbs). One would expect events to be described mostly in chronological order. This ordering represents a factual account of some sequence of events. There are a multitude of alternative orderings that might be used, among them a paragraph or two of the latest events (latest dates) followed by a chronological ordering of the remainder of the events. Another possible ordering is the reverse chronological ordering. Both of these orderings would present the latest developments first, on the assumption that they are the most relevant. This style of presenting the latest information first is frequently used in newspaper articles. Clearly, many different approaches to obtaining an ordering are available, including the sequence of particular words across the document (similar to Lapata (2003)'s

word-based features). However, these only incidentally capture chronological aspects, while mainly seizing information regarding topical sequences. I, therefore, explicitly use temporal orderings obtained via VerbOcean, which provides completely domain-independent temporal information.

Table 6.5 presents the results when using chronological and reverse chronological orderings. The first two rows check the ordering of events, while the latter two ensure that the corresponding sentences have a noun group in common, to increase the likelihood that two events are related.[5] The results clearly show that there is potential in the direct ordering of events, suggesting that coherence can to some degree be captured using simple temporal precedence orderings, without the need to model the topic sequences explicitly and dependent on the domain. In comparison to the accuracy achieved by other methods, however, the results are rather poor. Their combination with other features in later experiments will reveal whether the temporal features provide additional information and are useful in the overall model, or whether the features only represent information that is already captured by other features.

Temporal Ordering	**Accuracy**	
	Accidents	***Earthquakes***
Precedence Ordering	60.41	47.09
Reverse Ordering	39.59	52.61
Precedence with matching NG	**62.65**	**57.52**
Reverse with matching NG	37.35	42.48

Table 6.5: Results of Experiment 1 (Part 3). The impact of VerbOcean 'happens-before' temporal precedence relations on accuracy in the training datasets. "NG" denotes noun group.

Exploration 4. The last category of features investigated is the impact of longer range relations between sentences. In other words, can any benefit be obtained when considering the similarity of non-adjacent sentences? Under the assumptions of topically related sentences being grouped together, and topically similar sentences containing similar words, the sentences that are closest together should be most similar. Nevertheless, the size of the window of sentences that contain similar information has

[5] In case that VerbOcean does not identify any 'happens-before' relations, the instance counts both correct and incorrect with 0.5. As a consequence, the results reported here are aggregates for the whole dataset and not only based on the pairs in which VerbOcean determines relations.

yet to be determined. Table 6.6 presents the results of the relevant exploration. It provides the accuracy of the correctly ordered documents based on the similarity of sentences that are exactly n sentences away from each other, i.e., there are $n-1$ sentences between the sentence pairs being compared. The results are in line with those obtained thus far. As expected, for both datasets the similarity of directly adjacent sentences provides the best results. Yet, for the accident dataset, performance already drops steeply for a range of two sentences, while the drop is much lower in the earthquake dataset. This suggests that the topics in the documents of the earthquake dataset are typically expressed in between two and three sentences, while the topics tend to shift from sentence to sentence in the accident dataset. As indicated above, this is in line with the general layout of the documents.

Distance between Sentences (n)	**Accuracy**	
	Accidents	***Earthquakes***
1	**81.65**	**68.39**
2	59.93	64.11
3	51.73	53.02
4	49.88	51.86
5	47.83	52.17
6	51.00	50.54
7	53.48	51.27
8	53.05	50.93

Table 6.6: Results of Experiment 1 (Part 4). The impact of longer range relations on accuracy in the training datasets. The experiment uses noun and verb group chunk information for determining similarity.

Exploration 5. While the experiments presented thus far engineered a number of useful features, the following experiment integrates these features using SVMs[6] and compares the resulting system to other state-of-the-art systems. The results of this comparison, summarized in Table 6.7, show that the model with the features developed above performs reasonably well on the given synthetic datasets and, at least for the accident dataset, are in a similar league as the other systems. However, the model struggles with the topic-oriented earthquake dataset.

[6]The exact settings used for learning the model are: -z p -v 2 -t 1 -d 3

Combination	**Accuracy**	
	Accidents	***Earthquakes***
Chunk+Temp+WN+LongRange+	83.11	54.88
Chunk+Temp+WN+LongRange-	77.67	62.76
Chunk+Temp+WN-LongRange+	74.17	59.28
Chunk+Temp+WN-LongRange-	68.15	63.55
Chunk+Temp-WN+LongRange+	**86.88**	**63.83**
Chunk+Temp-WN+LongRange-	80.19	59.43
Chunk+Temp-WN-LongRange+	76.63	60.86
Chunk+Temp-WN-LongRange-	64.43	60.94
NG Similarity with Synonyms	85.90	63.55
Coreference+Syntax+Salience+	90.4	87.2
Coreference-Syntax+Salience+	89.9	83.0
HMM-based Content Models	75.8	88.0
Latent Semantic Analysis	87.3	81.0

Table 6.7: Results of Experiment 1 (Part 5). Comparison of the developed model with other state-of-the-art systems. "Chunk" denotes the chunking information (i.e., noun, verb group, and head information); "Temp" denotes temporal information (i.e., the use of relative temporal position obtained using VerbOcean); "WN" denotes WordNet information (i.e., synonym, hypernym, etc. relations), and "LongRange" denotes the use of similarities between sentences that are not directly adjacent. +/- following the information type denotes whether or not that type of information was available to the learning algorithm. The results of the following systems are reproduced from Barzilay and Lapata (2008): Coreference+Syntax+Salience+ and Coreference-Syntax+Salience+ are two versions of Barzilay and Lapata (2008)'s model, "HMM-based Content Models" represents Barzilay and Lee (2004)'s approach, and "Latent Semantic Analysis" is Barzilay and Lapata (2008)'s implementation of Foltz et al. (1998). The Coreference+Syntax+Salience+ model uses co-reference resolution (based on the original sentence ordering for resolution in alternative renderings of the texts), the syntactic position of the entities (subject, object, other), and salience (distinguishing between salient and non-salient entities given the entities' frequencies).

While the results indicate that my model performs significantly worse than the best Barzilay and Lapata (2008) model, when disregarding co-reference (Coreference-Syntax+Salience+; owing to its gold standard character), the differences in the earthquake domain are notably smaller. One major difference between Barzilay and Lapata (2008) and my shallow model is that my model only considers similarity between sentence pairs. Barzilay and Lapata, in turn, utilize sequences across multiple sentences, which to a large degree capture coherence between topics, i.e., sentences on the same topic tend to appear close to each other. The results also illustrate the differences in performance for approaches based on very local coherence measures (my approach), larger local context (Barzilay and Lapata for entity sequences exceeding two), and global views of coherence (topical view of coherence in the HMM-based Content Models). Yet, notwithstanding its shortcomings, an essential advantage of my approach is that it is quite shallow and could easily be incorporated into the sentence selection process as the features can be computed once for the sentence pairs and then looked up for a particular ordering as opposed to a re-computation of all features.

6.4.3 Experiment 2: The Application of Sentence Ordering to Automatically Generated Summaries

As stated, the ultimate goal of the models considered in this chapter is the application of sentence ordering to automatically generated summaries. In this regard, there is one major difference between coherence as studied in Experiment 1 and coherence in the context of automatic summarization. Namely, the topics of the documents are unknown at the time of training for newswire summarization systems. As a result, model performance on *out-of-domain* texts is a crucial attribute for good performance in the realm of summarization.

Experiment 2 seeks to evaluate how well my model performs in such cases. To this end, I perform two sets of tests. First, following the strategy of Barzilay and Lapata (2008), I cross-train the models between the accident and earthquake datasets to determine system performance in unseen domains. Second, I use the dataset based on the DUC2005 reference summaries to investigate whether my model's performance on unseen topics reaches a plateau after training on a particular number of different topics.

The results of cross-training between the accident and earthquake datasets are summarized in Table 6.8. They show that my shallow model comprising temporal re-

Chunk+Temporal+WordNet+LongRange+

Train \ Test	Earthquakes	Accidents
Earthquakes	62.74	63.6
Accidents	56.37	78.4

Chunk+Temporal-WordNet+LongRange+

Train \ Test	Earthquakes	Accidents
Earthquakes	63.83	**86.63**
Accidents	**64.19**	86.88

Coreference+Syntax+Salience+

Train \ Test	Earthquakes	Accidents
Earthquakes	87.3	67.0
Accidents	69.7	90.4

HMM-based Content Models

Train \ Test	Earthquakes	Accidents
Earthquakes	88.0	31.7
Accidents	60.3	75.8

Table 6.8: Results of Experiment 2 (Part 1). Cross-training between the accident and earthquake datasets. The results for the Coreference+Syntax+Salience+ and HMM-Based Content Models reported in the third and fourth table are reproduced from Barzilay and Lapata (2008).

lations from VerbOcean (first table) performs considerably worse than the syntactically and theoretically motivated model by Barzilay and Lapata (2008) (third table). The model without temporal relations (second table), however, performs rather well. While the model trained on the earthquake dataset performs worse than the comparable model by Barzilay and Lapata (2008), my model trained on the accident dataset performs better, and by a considerable margin. My model also performs better than the HMM-based content models (fourth table; Barzilay and Lee (2004)), particularly for the cross-trained sections.

A comparison between the first and second table suggests once again that using the temporal relations from VerbOcean does not ultimately promote performance. In fact, inclusion of temporal features provides significantly inferior results on cross-trained data. The results of the model using all features except temporal ones (reported in the second table) indicates that the model developed in this chapter provides excellent generalization to unseen datasets. The performance is nearly as good on cross-trained data as on the original data. Comparing these results to Barzilay and Lapata (2008) and Barzilay and Lee (2004) shows that their models fall drastically in performance when using cross-training. This, in turn, indicates that shallow features can be used successfully to develop a sentence ordering system that provides good generalization capabilities to unseen data genres, as is crucial for sentence ordering in the context of automatic summarization.

Table 6.9 presents the results when cross-training the system on different *numbers* of topics. While the approach is somewhat biased due to the increasing amounts of training data, the results nonetheless provide some insight into the problem of out-of-domain texts since good performance requires an abstraction from specific topics. Note that, for the problem of summarization, the amount of training data available is usually sizeable. As this experiment only uses some data – from one specific year – the reported results therefore, in effect, represent a lower bound on performance.

The results are remarkable. Considering the amount of training data, which is less than a quarter of the training data available for cross-training in the experiments underlying Table 6.8, the system already achieves a respectable score of 72% on new topics. This indicates that the goal of constructing a sentence ordering model that performs well irrespective of the topic of the document has been achieved. It is to be expected that, as the amount of training data increases – be it in the form of more topics or more examples within the topics – performance will increase even further. An interesting, more general point in this regard is that the model does not require much

Topics for Training	Training Examples	Accuracy
2	84	68.48
4	168	60.05
6	252	62.28
8	336	63.27
10	463	**72.28**

Table 6.9: Results of Experiment 2 (Part 2). Accuracy on five test topics from DUC2005 with respect to the number of topics used for training using the Chunk+Temporal-WordNet+LongRange+ model.

training data. The problems from over-conforming to a particular domain as illustrated by the results in the first experiment of this section are thus easily avoided by using a number of different topics.

6.4.4 Experiment 3: Sentence Ordering Based on the Sentences Selected by my Summarization System (Chapter 5)

Both of the foregoing experiments investigated the model and its generalization capabilities on the basis of synthetic datasets. Ultimately, however, the model is to be used as an integral part of an automatic summarization system. While I do not provide a full evaluation of the model within a human study (due to time constraints), this section assesses the performance of the sentence ordering approach in the form of a visual inspection of textual output derived in Chapter 5. In particular, for the sample information selected as summary content in Figures 5.3, 5.5, and 5.7, Figures 6.5, 6.6, and 6.7 provide the results when ordering the information as per the model developed in this chapter. Since frequently more than one sentence ordering is (equally) plausible according to the rules set out by the procedure, such cases are resolved by selecting the ordering that most resembles the order in which the sentences were selected, namely, the most important information is presented first. The background colors in the figures denote matches identified between the individual sentences.

The first example, presented in Figure 6.5, shows that determining a good ordering of the sentences is quite difficult. Examining the colored boxes, it is easy to see that several of the relevant terms, e.g., “India” and “Pakistan,” occur in all four sentences. As such, the main distinguishing groups are “Kashmir,” “Kashmiri,” and “rival,” that

is, these are groups that only occur in a subset and can thus be grouped together, thereby imposing an ordering on the sentences. Although not identified as similar by the WordNet relations utilized as part of the sentence ordering approach, the former two groups are very much alike. As a result of these matches, the first and second sentence should be directly adjacent to each other, as should the second and third, and the third and the fourth. Yet, there are two orderings fulfilling these constraints, the one presented in the figure and the reverse order of the one presented in the figure. The order in the figure is preferred because of the order of selection of the sentences (based on Chapter 5.

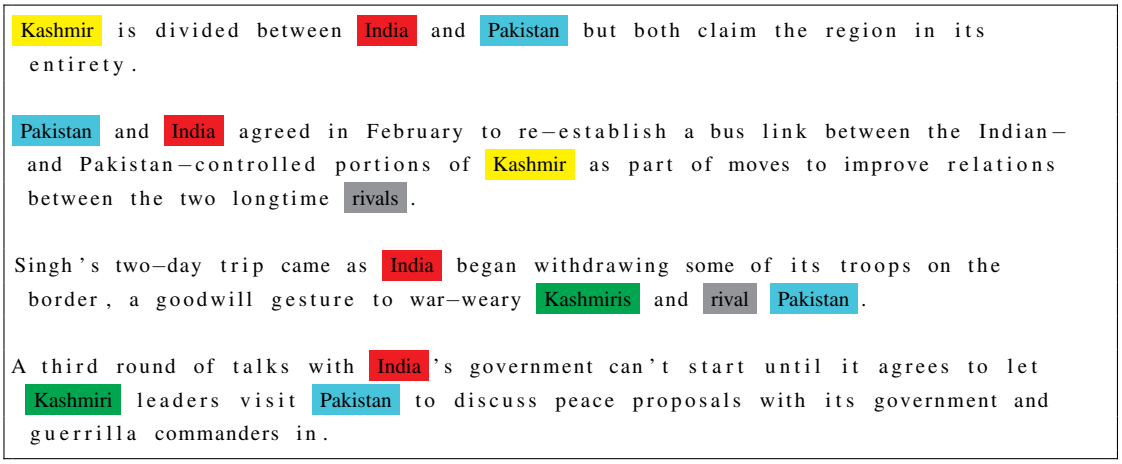

Figure 6.5: The ordering of the sentences in Figure 5.3 according to the ordering approach in this chapter.

The second example, displayed in Figure 6.6, is similar to the first one, but (also) illustrates that the ordering of sentences is important, as the order selected by the approach presented in this chapter mis-orders the information, resulting in misleading information. Based on the matches identified by the colored boxes, the first and second sentence should occur next to each other, as should the second and third, and third and fourth. Again, there are two possible orderings, with the preferred being the one presented in the figure. However, if one examines the original version of the last sentence (*cf.* Figure 5.5), one notices that the IRA declared a ceasefire in 1998, before the Good Friday peace deal came into force. The order of sentences and the truncated nature of the last sentence, on the other hand, give the impression that the IRA declared a ceasefire at the end of the period covered in the document set, a fact which is wrong!

The final example in Figure 6.7 illustrates an instance that allows for numerous orderings of the sentences, as the overlap between the sentences is minimal. The fact

The 1998 Good Friday peace deal paved the way for a Protestant–Catholic power–sharing assembly, but that was suspended more than two years ago amid allegations of IRA espionage.

Protestant factions in Northern Ireland are adamant that there can be no political progress toward a lasting peace settlement without a move by the IRA to end all criminal and paramilitary activity.

Ahern said he was hopeful that the IRA would issue a declaration within the next few months spelling out whether it will disarm and cease all threatening activities.

The IRA declared a ceasefire in its campaign to end British rule.

Figure 6.6: The ordering of the sentences in Figure 5.5 according to the ordering approach in this chapter.

that the last sentence does not have any overlaps means that it can be placed at any position in the order so long as it does not interrupt the relations between the other sentences. In this case, this means either before or after the other three sentences. Similarly, the second sentence should be before or after the first and third, while no ordering is defined for the first and third sentences themselves. As a result, there is no guarantee of ordering that sentence in any way that is coherent, as there is no measure of determining a position at which the sentence should occur. Once again, the ordering presented in the figure is the order preferred on the basis of the order of sentence selection. This example very clearly shows that the use of anaphora resolution on the original documents and the use of that information in the sentence ordering stage would result in more potential matches between the different sentences. At the very least, using anaphora resolution, a link between the last three sentences can be identified, thus determining the position of the last sentence based on the quality of the ordering as opposed to the importance of the sentence in the sentence selection stage.

Overall, the examples in this section reveal that the ordering of the sentences achieved via the proposed algorithm is mostly quite reasonable, though this is largely due to the preferred order as given by the sentence selection process. Otherwise, chronological and reverse chronological order would be equally likely. Even so, the order in the examples is satisfactory, despite the fact that it is not always in line with the actual chronological order. It would thus appear that the incorporation of the date of some event or, that being absent, the date of publication of the content units used to select the sentences would provide an additional useful means of ordering the sentences.

Sandra Day O'Connor, the first woman appointed to the US Supreme Court and a frequent swing vote, announced her retirement Friday, setting up a fierce political showdown for her seat.

US President George W. Bush on Wednesday chastised critics of Attorney General Alberto Gonzales as a potential candidate for the recently-vacated seat in the Supreme Court.

He also promised to "be deliberate and thorough" in making the choice and will consult with the Senate which will confirm his nominee.

"He said he's going to pick a strong conservative, and I think that's going to cause a battle no matter what,".

Figure 6.7: The ordering of the sentences in Figure 5.7 according to the ordering approach in this chapter.

6.5 Discussion

The purpose of this chapter was to develop an appropriate method for the last step involved in the generation of natural language summaries – sentence ordering. Indeed, the experiments support the notion that my shallow, topic-independent, unlexicalized, local models of coherence achieve respectable performance in sentence ordering tasks when compared to other state-of-the-art systems, especially if the domain is unknown, as is the case for automatic summarization systems. In particular, the present work established the usefulness of WordNet relations for the task, which differ with respect to the specific relations employed, and chunker information, which provides small improvements compared to unchunked texts. Last but not least, using both cross-trained data on the Barzilay and Lapata (2008) earthquake and accident datasets as well as a topic-dependent training scenario on DUC2005 data, I showed that my model generalizes at a good rate to unseen topic domains. The sample output for the sentence ordering model when applied to the output of the sentence selection algorithm developed in Chapter 5 also indicates that the approach provides satisfactory orderings of the selected information.

More generally, the present work shows that relatively shallow approaches provide sufficient topic-independence compared to other approaches to be useful in generic summarization, although up to now I only compared different orderings of the same sentences. In future work, I intend to extend my model in order to judge the relative coherence of texts containing different sentences (similar to the Barzilay and Lapata

(2008) experiment on DUC data). This is likely to be useful both with respect to the evaluation of the coherence of summaries as well as the generation of more coherent summaries by integrating coherence constraints into the sentence selection stage of the summarization process.

Chapter 7

Conclusion and Future Work

This thesis on automatic summarization had three main objectives: (1) The development of an automatic evaluation method; (2) drawing on the ideas used to construct the evaluation method, the development of an automatic multi-document summarization system; and (3) the development of a method for sentence ordering based on text coherence. While many of the issues and questions motivating the present work have been resolved, there remains substantial scope for future work. The purpose of this chapter is to summarize the main contributions of this thesis and indicate several avenues for prospective projects.

Contributions

Despite their limitations, the methods and algorithms developed as part of this thesis have expanded the field of automatic summarization in multiple respects. For one, the fully automatic evaluation system for automatically generated summaries based on the concepts underlying the manual, state-of-the-art Pyramid evaluation method outperforms the best current, universally accepted, automatic evaluation method – ROUGE. As such, it provides a means for more accurate system development without significant expenditure on manual labor.

The work on the automatic system to generate summaries moreover shows that various aspects of the information derived via the (original) manual Pyramid scheme can be exploited beyond the evaluation context. In particular, the annotation of the Pyramid scheme can be used to obtain information regarding the semantic similarity of the information contained in the source documents as well as its relative importance. In addition, it provides a channel to develop an approach that selects variable-sized informational units.

One of the main contributions of these efforts to summarization research is the view of the set of source documents from a semantic perspective, that allows for the aggregation of similar content into non-atomic, variable-sized units of information. A key advantage of this approach is its abstraction from the syntactic structure of the source documents and choice of words for the summary text. From this perspective, the proposed content-unit-based summarization approach provides a step towards the actual generation of summaries as opposed to the extraction of (simplified) sentences from the source documents. In other words, the content units enable one to determine accurately the exact content that needs to be (re)generated into a summary.

As a final point, the proposed method for sentence ordering reveals that a number of features at the shallow end of sentence ordering can successfully be applied to the

summarization context, as shown by the success of the subset selected for the proposed summarization system. Although many current approaches to sentence ordering exploit deep syntactic features, their performance decreases rapidly on out-of-domain documents. My shallow approach, in turn, (still) performs quite well in such scenarios. Likewise, applying the approach to the sentence selected for summarization results in summaries that are easily readable and mostly maintain the "correct" order of information. In short, I present a sentence ordering approach that generalizes well to unseen domains.

Future Work

The most promising routes with respect to enhancing some of the proposed techniques and/or expanding the present work in informative directions can be classified into three categories.

Component Interaction. One of the most interesting aspects for future work is the integration of the information selection and information ordering approaches into a single, consistent summarization system. The present work uses the most common approach for generating the final summary: the system first selects the most important information from the collection of source documents (c.f., Chapter 5) and, in a second processing step, orders it as intelligibly as possible (c.f., Chapter 6). In other words, information ordering is a *post*-processing step to information selection. Yet, as indicated in Chapter 2, information is frequently selected in order to create a coherent summary as opposed to selection purely on the basis of informational content. As such, the use of joint models that embrace both aspects at the same time is bound to create a better automatic summarization system, as the system would be able to exploit (some of) the additional information used by human summarizers. Examples of the successful implementation of such joint models in natural language processing are the joint modeling of sentence extraction and sentence compression in summarization (Martins and Smith, 2009), joint modeling of co-reference resolution and named-entity classification (Denis and Baldridge, 2009), joint modeling of syntactic and semantic labeling (Lluís and Màrquez, 2008), and joint modeling of tagging and parsing (Johnson, 2001).

Improvement of Individual Components. Other instructive aspects for future work relate to the improvement of individual components of the various algorithms, most notably the matching procedure for individual concepts or template instantiations and the aggregation of individual informational units into content units. In the context of the former, for instance, use of more complex syntactic templates in conjunction

with the exploitation of further knowledge resources – such as paraphrase information (DIRT; Lin and Pantel, 2001) – is likely to result in (even) better performance.

As regards the identification of content units, while the present clustering and aggregation algorithms do perform well, further investigation into the features used in the algorithms as well as the structure of the algorithm itself may improve performance even more. For example, besides syntactic and proximity constraints, the algorithm might also make use of such features as graph centrality and discourse-related aspects to facilitate the determination of cluster association. From an algorithmic perspective, a wide variety of clustering approaches can be used to determine the ultimate cluster composition, including fuzzy clustering using expectation maximization algorithms (Dempster et al., 1977). It might likewise be useful to investigate whether the use of a more complete set of WordNet relations (e.g., troponyms, entailment, holonyms) would be beneficial. Their use should, on the one hand, result in more accurate detection of similarity. On the other hand, however, use of all of the relevant relations might result in an explosion of possible matches and therefore give rise to poorer similarity detection overall.

Moving beyond advancements for the selection of content units themselves, note that the proposed summarization system only generates generic summaries in the sense that a set of documents is summarized irrespective of any constraints relating to the (required) content of the summaries. Most contemporary summarization systems, however, support the generation of summaries based on a topic statement, which defines what content is interesting for the summary. Recent research for the TAC "Summarization" track, for instance, is concerned with the generation of so-called "update" summaries, i.e., summaries that condense a selection of documents given the assumption that the reader is already familiar with a (different) set of related documents. My summary generation system based on Pyramid-style content units should, in principle, (also) perform well in that context because it considers the semantic content of a textual unit explicitly, although the relevant investigations(s) are still outstanding. It should, correspondingly, be relatively straightforward to determine which content units occur in both sets of documents and to adjust the summary content accordingly. A related role for the content units used to generate the summary is to replace words or phrases on the basis of the matches found in the content units. Doing so might improve the ensuing summaries in at least two ways: (1) Sentences in the summary might be shortened based on the information from the content units, and (2) co-references might be resolved such that the summary is more readable by replacing a co-reference with the

actual entity.

The proposed sentence ordering procedure also leaves scope for improvement. For one, it could use more of the syntactic and tense information provided by LT-TTT2. For example, considering the differences in performance when using lemmas versus surface forms, using lemmas and syntactic information such as tense, person, and grammatical voice might achieve the best of both worlds by facilitating more detailed matching of the syntactic information and, at the same time, providing more generalization of the semantic meaning of verbs. A second area for development regards the incorporation of sentence ordering into the summarization process. As is, the sentence ordering procedure does not utilize any information obtained from the source documents. However, using such information – e.g., the most likely order of the content units based on the order of content units in the source documents, or the temporal order of the content units based on the date of publication of the articles – is likely to be a good (and quite simple) way of enhancing the ordering of information in the summaries generated by way of the proposed automatic summarization system.

On a more general note, all of the components of the proposed system can potentially be improved by using anaphora resolution, e.g., BART (Versley et al., 2008). Doing so, similarity between constituents could be determined more accurately, which would result in improved summarization evaluation and summary generation. To be precise, anaphora resolution can help in two central respects: (1) it facilitates the resolution of pronouns to *one* specific antecedent as opposed to the multiple antecedents employed in the current implementation, and (2) it enables the grouping of noun groups/entities that describe the same entity. Aside from its usefulness in the content unit creation/matching processes, anaphora resolution might also be useful when determining the ordering of sentences. The current implementation does not consider anaphora resolution at all because the antecedents cannot be determined accurately in all sentence ordering settings. Yet, for sentence ordering in the context of summarization, the antecedents determined on the original documents can in principle be passed on to the sentence ordering component and be utilized in this context. This would result in more links between the sentences, which in turn should result in more coherent sentence orderings.

Translating the Techniques to Different Contexts. For the purposes of this thesis, the syntactic templates and relationships between the constituents of different templates were manually defined, and similarities between the different constituents were identified on the basis of their (syntactic) structures. Unfortunately, though not unex-

pectedly, the syntactic templates did not capture all possible (relevant) syntactic variations, and it would be difficult to enumerate the possibilities manually. Yet, combining the syntactic templates with human judgments regarding semantic content (in the form of SCUs and PeerSCUs), it might be possible to induce further frequently used syntactic structures (exploiting frequency to avoid the identification of each and every syntactic relation provided by the parser). In its essence, the suggested procedure performs a similar task as DIRT (Lin and Pantel, 2001), but utilizes the annotated information available and should thus obtain much more accurate information, which – at a later stage – might constitute seed information based on which a DIRT-like approach could be used on unsupervised data.

Apart from the ability to learn syntactic structures, this approach could also be employed to learn semantic similarities between words. In many of my examples I remarked on relationships that could not be identified based on WordNet. Use of the Pyramid annotation along with identified syntactic templates should allow for the determination of similarities between words with high accuracy. In one of the examples, information sharing performed this task. Yet, based on such inferences, it would be feasible to develop a highly accurate automatically generated knowledge resource that complements the information provided by WordNet. In general, the approach would work well in combination with the induction of similarities between syntactic structures. Using both in an iterative manner, one could employ the improvements made in one of the two areas to enhance the accuracy of the other. What is more, although both of the foregoing approaches would provide useful information in and of themselves, the information obtained by these methods could, in turn, be used to expand the proposed automatic Pyramid-style evaluation and summarization systems (that is, they improve the overall system by providing information that can improve the individual components of the evaluation framework).

Bibliography

Baldewein, U., Erk, K., Pado, S., and Prescher, D. (2004). Semantic role labeling with similarity based generalization using em-base clustering. In *Senseval 2004*.

Barzilay, R. (2003). *Information fusion for multidocument summarization: paraphrasing and generation*. PhD thesis, Columbia University.

Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization.

Barzilay, R., Elhadad, N., and Mckeown, K. R. (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:2002.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Comput. Linguist.*, 34:1–34.

Barzilay, R. and Lee, L. (2004). Catching the drift: probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71.

Bollegala, D., Okazaki, N., and Ishizuka, M. (2005). A machine learning approach to sentence ordering for multidocument summarization and its evaluation. In Dale, R., Wong, K.-F., Su, J., and Kwong, O. Y., editors, *Natural Language Processing Ű IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 624–635. Springer Berlin-Heidelberg.

Bollegala, D., Okazaki, N., and Ishizuka, M. (2006). A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of ACL-44*.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM.

Bosma, W. (2009). Contextual salience in query-based summarization. In *Proceedings of the International Conference RANLP-2009*, pages 39–44, Borovets, Bulgaria. Association for Computational Linguistics.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80, Morristown, NJ, USA. Association for Computational Linguistics.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*.

Chowdary, C. and Kumar, P. (2009). Esum: An efficient system for query-specific multi-document summarization. In Boughanem, M., Berrut, C., Mothe, J., and Soule-Dupuy, C., editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 724–728. Springer Berlin / Heidelberg.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the EACL 2003*.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Conroy, J., Schlesinger, J., and Stewart, J. G. (2005). Classy query-based multi-document summarization. In *Proceedings of DUC 2005*.

Conroy, J. M. and O'Leary, D. P. (2001). Text summarization via hidden markov models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407, New York, NY, USA. ACM.

Conroy, J. M., Schlesinger, J. D., and O'Leary, D. P. (2009). Classy 2009: Summarization and metrics. In *Text Analysis Conference 2009*.

Copeck, T., Inkpen, D., Kazantseva, A., Kennedy, A., Kipp, D., Nastase, V., and Szpakowicz, S. (2006). Leveraging duc. In *Document Understanding Workshop (DUC 2006)*.

Dang, H. (2005). Overview of duc 2005.

Daumé III, H., Langford, J., and Marcu, D. (2009). Search-based structured prediction. *Machine Learning Journal*, 75(3):297–325.

Daumé III, H. and Marcu., D. (2005). Bayesian multi-document summarization at mse. In *Proceedings of MSE 2005*.

Daumé III, H. and Marcu, D. (2005a). Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*.

Daumé III, H. and Marcu, D. (2005b). Bayesian summarization at duc and a suggestion for extrinsic evaluation. In *Document Understanding Conference 2005*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):pp. 364–366.

DeJong, G. (1982). An overview of the frump system. In Lehnert, W. G. and Ringle, M. H., editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38. Series B.

Denis, P. and Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. In *Procesamiento del Lenguaje Natural 42*.

Dom, B. E. (2001). An information-theoretic external cluster validation measure. *Journal of American statistical Association*, 78:553–569.

Edmundson, H. P. (1969). New methods in automatic extracting. *J. ACM*, 16(2):264–285.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–479.

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press.

Fisher, D. H. and Pazzani, M. J. (1991). Computational models of concept learning. In Fisher, D. H., Pazzani, M. J., and Langley, P., editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann, San Mateo, CA:.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). Textual coherence using latent semantic analysis. *Discourse Processes*, 25:285–307.

Geiss, J. (2009). Creating a gold standard for sentence clustering in multi-document summarization. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages pp 96–104, Suntec, Singapore. ACL and AFNLP.

Giannakopoulos, G. and Karkaletsis, V. (2009). N-gram graphs: Representing documents and document sets in summary system evaluation. In *Text Analysis Conference 2009*.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288.

Gong, Y. and Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA. ACM.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Grover, C. and Tobin, R. (2006). Rule-based chunking and reusability. In *Proceedings of LREC 2006*.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Harnly, A., Nenkova, A., Passonneau, R., and Rambow, O. (2005). Automation of summary evaluation by the pyramid method. In *In Proceedings of the Conference of Recent Advances in Natural Language Processing (RANLP*, page 226.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., and McKeown, K. R. (2001). Simfinder: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*, pages pp. 41–49. ACL.

He, T., Li, F., Shao, W., Chen, J., and Ma, L. (2008). A new feature-fusion sentence selecting strategy for query-focused multi-document summarization. In *Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology*, pages 81–86, Washington, DC, USA. IEEE Computer Society.

Heidorn, G. (2000). Intelligent writing assistance. In R.Dale, H. and H.Somers, editors, *A Handbook of Natural Langauge Processing: Techniques and Applications for the Processing of Language as Text*. New York: Marcel Dekker.

Hess, A. and Kushmerick, N. (2003). Automatically attaching semantic metadata to web services. In *Proceedings of the 2nd International Semantic Web Conference*, Florida, USA.

Hirao, T., Isozaki, H., Maeda, E., and Matsumoto, Y. (2002). Extracting important sentences with support vector machines. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Hovy, E., Lin, C.-Y., and Zhou, L. (2005). Evaluating duc 2005 using basic elements. In *Document Understanding Conference 2005*.

Hovy, E., yew Lin, C., Zhou, L., and Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC*.

Iyer, R. and Ostendorf, M. (1996). Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models. In *IEEE Transactions on Speech and Audio Processing*, pages 236–239.

Ji, P. D. and Pulman, S. (2006). Sentence ordering with manifold-based classification in multi-document summarization. In *Proceedings of EMNLP 2006*.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pages 310–315, Morristown, NJ, USA. Association for Computational Linguistics.

Joachims, T. (2002). Evaluating retrieval performance using clickthrough data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*.

Johnson, M. (2001). Joint and conditional estimation of tagging and parsing models. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 322–329, Morristown, NJ, USA. Association for Computational Linguistics.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Katragadda, R. (2009). On alternative automated content evaluation measures. In *Text Analysis Conference 2009*.

Kittredge, R., Korelsky, T., and Rambow, R. (1991). On the need for domain communication knowledge. *Computational Intelligence*, 7:305–314.

Klein, D. (2005). *The unsupervised learning of natural language structure*. PhD thesis, Stanford University.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91 – 107.

Kumar, S. and Kumar, A. (2009). Iit kharagpur at tac 2009: Statistical and nugget-based model for automatic summary evaluation. In *Text Analysis Conference 2009*.

Kuno, S. (1972). Functional sentence perspective: a case study from japanese and english. *Linguistic Inquiry*, 3:269–320.

Kupiec, J., Pedersen, J. O., and Chen, F. (1995). A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 68–73.

Lalitha Devi, S., Kuppan, S., Venkataswamy, K., and Rao, P. R. (2009). Identification of similar documents using coherent chunks. In *DAARC '09: Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications*, pages 54–68, Berlin, Heidelberg. Springer-Verlag.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proc. of ACL 2003*.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *KDD 1999*.

Leskovec, J., Grobelnik, M., and Milic-Frayling, N. (2004). Learning sub-structures of document semantic graphs for document summarization. In *LinkKDD 2004*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *In Proceedings of HLT/NAACL 2003*.

Lin, D. (1998). Dependency-based evaluation of minipar. In *Proc. Workshop on the Evaluation of Parsing Systems*, Granada.

Lin, D. and Pantel, P. (2001). Dirt - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328.

Lluís, X. and Màrquez, L. (2008). A joint model for parsing syntactic and semantic dependencies. In *CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 188–192, Morristown, NJ, USA. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts'. *IBM Journal of Research and Development*, 2:159–165.

Ma, L., He, T., Li, F., Gui, Z., and Chen, J. (2008). Query-focused multi-document summarization using keyword extraction. In *Computer Science and Software Engineering, 2008 International Conference on*, volume 1, pages 20 –23.

Ma, Y.-F., Lu, L., Zhang, H.-J., and Li, M. (2002). A user attention model for video summarization. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, New York, NY, USA. ACM.

Madnani, N., Passonneau, R., Ayan, N. F., Conroy, J. M., Dorr, B. J., Klavans, J. L., O'Leary, D. P., and Schlesinger, J. D. (2007). Measuring variability in sentence ordering for news summarization. In *Proceedings of the 11th European workshop on Natural Language Generation*.

Manning, C., Raghavan, P., and Schuetze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marcu, D. (1997). From local to global coherence: a bottom-up approach to text planning. In *in Proceedings of AAAI-97, American Association for Artificial Intelligence*.

Marcu, D. (1998). Improving summarization through rhetorical parsing tuning.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pages 123–136. The MIT Press.

Marcu, D. and Gerber, L. (2001). An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, Pittsburgh, PA.

Martins, A. F. T. and Smith, N. A. (2009). Summarization with a joint model for sentence extraction and compression. In *ILP '09: Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.

McKeown, K. R. (1985). *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press.

McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. (1999). Towards multidocument summarization by reformulation: Progress and prospects. In *IN PROCEEDINGS OF AAAI-99*, pages 453–460.

Meila, M. (2007). Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98(5):pp. 873–895.

Miike, S., Itoh, E., Ono, K., and Sumita, K. (1994.). A full text retrieval system. In *SIGIR '94*.

Minnen, G., Carroll, J., and Pearce, D. (2000). Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*.

Mirkin, B. G. (1996). *Mathematical classification and clustering*. Kluwer Academic Press.

Miyao, Y. and Tsujii, J. (2008). Feature forest models for probabilistic hpsg parsing. *Comput. Linguist.*, 34(1):35–80.

Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *in Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. Physica-Verlag, Vienna.

Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 763–772, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nenkova, A. and Passanneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarisation. Technical report, MSR-TR-2005-101. Microsoft Research.

Nenkova, A., Vanderwende, L., and McKeown, K. (2006). A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR*

conference on Research and development in information retrieval, pages 573–580, New York, NY, USA. ACM.

Newman, E., Doran, W., Stokes, N., Carthy, J., and Dunnion, J. (2004). Comparing redundancy removal techniques for multi-document summarisation. In *Proceedings of STAIRS*.

Olson, C. F. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing*, 21:1313–1325.

Olson, D. L. and Delen, D. (2008). *Advanced Data Mining Techniques*. Springer.

Ouyang, Y., Li, S., and Li, W. (2007). Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 79–86, New York, NY, USA. ACM.

Ouyang, Y. and Li, W. (2009). Polyu at tac 2009. In *Text Analysis Conference 2009*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *In Proceedings of the 40th Annual Meeting of the ACL*.

Passonneau, R. and Nenkova, A. (2003). Evaluating content selection in human- or machine-generated summaries: The pyramid method. Technical report, Technical report CUCS-025-03, Columbia University.

Passonneau, R., Nenkova, A., McKeown, K., and Sigleman, S. (2005). Applying the pyramid method in duc 2005. In *Document Understanding Conference (DUC'05)*.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1988). *Numerical Recipies in C: The art of Scientific Programming*. Cambridge University Press.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., and Zhang, Z. (2004). MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal.

Radev, D., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization 2000*.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Celebi, A., Liu, D., and Drabek, E. (2003). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Radev, D. R. and McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Comput. Linguist.*, 24(3):470–500.

Rambow, O. (1990). Domain communication knowledge. In *Proceedings of the Fifth International Workshop on Natural Language Generation*.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *American Statistical Association Journal*, 66(336):846–850.

Reichart, R. and Rappoport, A. (2009). The nvi clustering evaluation measure. In *Proceedings of CoNLL 2009*.

Reiter, E. and Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: a conditional entropy-based external clustering measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages pp.410–420.

Sagae, K., Miyao, Y., and Tsujii, J. (2007). Hpsg parsing with shallow dependency constraints. In *Proc. ACL 2007*, pages 624–631.

Saggion, H., Radev, D. R., Teufel, S., and Lam, W. (2002). Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *COLING*.

Shen, R., Nahnsen, T., Grover, C., and Klein, E. (2008). Recognising textual entailment focusing on non-entailing text and hypothesis. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

Shin, S.-I. and Choi, K.-S. (2004). Automatic word sense clustering using collocation for sense adaptation. In *Proceedings of the Second Global WordNet Conference*.

Siddharthan, A., Nenkova, A., and McKeown., K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of COLING 2004*.

Sparck Jones, K. (1998). Automatic summarising: factors and directions. In Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization.* MIT Press: Cambridge, MA, USA.

Sparck Jones, K. and Galliers, J. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York: Springer.

Spath, H. (1985). *The Cluster Dissection and Analysis Theory FORTRAN Programs Examples*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R., and Poesio, M. (2009). Wb-jrc-ut's participation in tac 2009: Update summarization and aesop tasks. In *Text Analysis Conference 2009*.

Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Articles*. PhD thesis, University of Edinburgh.

Teufel, S. and van Halteren, H. (2003). Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT/NAACL 2003 on Text summarization workshop*.

Thione, G. L., van den Berg, M., Polanyi, L., and Culy, C. (2004). Hybrid text summarization: Combining external relevance measures with structural analysis. In *ACL 2004*.

Toutanova, K., Markova, P., and Manning, C. D. (2004). The leaf projection path view of parse trees: Exploring string kernels for hpsg parse selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Tratz, S. and Hovy, E. (2009). Bewt-e for tac 2009's aesop task. In *Text Analysis Conference 2009*.

Tucker, R. (1999). *Automatic summarising and the CLASP system*. PhD thesis, University of Cambridge.

Tucker, R. and Sparck-Jones, K. (2005). Between shallow and deep: an experiment in automatic summarising. Technical report, Technical Report UCAM-CL-TR-631. University of Cambridge.

van Dongen, S. (2000). Performance criteria for graph clustering and markov cluster experiments. Technical report, CWI, Amsterdam.

van Halteren, H. and Teufel, S. (2003). Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

van Halteren, H. and Teufel, S. (2004). Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of EMNLP 2004*.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth, 2nd edition.

Vanderwende, L., Banko, M., and Menezes, A. (2004). Event-centric summary generation. In *Document Understanding Conference 2004*.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, 43(6):1606–1618.

Versley, Y., Moschitti, A., Poesio, M., and Yang, X. (2008). Coreference systems based on kernels methods. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 961–968, Morristown, NJ, USA. Association for Computational Linguistics.

Vlachos, A., Korhonen, A., and Ghahramani, Z. (2009). Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the EACL workshop on Geometrical Models of Natural Language Semantics*.

Wang, D., Li, T., Zhu, S., and Ding, C. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *SIGIR'08:*

Proceedings of the 31st international ACM SIGIR conference on research and development in information retrieval, pages 307–314, New York, NY. ACM.

Wormeli, R. (2005). *Summarization in any subject: 50 techniques to improve student learning*. Association of Supervision and Curriculum Development, Alexandria, VA.

Yakushiji, A., Miyao, Y., Tateisi, Y., and A, A. J. T. (2005). 2005. biomedical information extraction with predicateargument structure patterns. In *In Proceedings of The Eleventh Annual Meeting of The Association for Natural Language Processing*, pages 60–69.

Yarowsky, D. (2000). Word sense disambiguation. In Dale, R., Moisl, H., and Somers, H., editors, *The Handbook of Natural Language Processing*, pages 629–654. Marcel Dekker.

Zeng, Y., Tang, J., Garcia-Frias, J., and Gao, G. (2002). An adaptive meta-clustering approach: combining the information from different clustering results. In *IEEE Computer Society Bioinformatics Conference (CSB 2002)*.

Zhao, Y. and Karypis, G. (2001). Criterion functions for document clustering: experiments and analysis. Technical report, Department of Computer Science, University of Minnesota.

Appendix A

Reference Summaries for Document Collection D324

This appendix contains a representative set of reference summaries, D324, reproduced from DUC2005 (annotated with Pyramid annotations). D324 regards British and Argentinian relations over the Falkland Islands. The alpha-numeric title completely identifies the source dataset (D324), the length (250 words), the annotator (Individual E), and the composer of the reference summary (Individual D).

D324.M.250.E.D

Although there are still remaining issues regarding the Falkland Islands, relations between Argentina and Great Britain are growing continuously cooperative. President Carlos Menem of Argentina who took office in 1989 has adopted a pro-western foreign policy and free market economics. This has paved the way to economic and diplomatic relations with Great Britain. Argentina has sought British assistance on privatization of the oil and gas industries and the development of oil resources. Economic disputes still exist over the fishing industry in the waters around the Falkland Islands, particularly the lucrative illex squid fishing. In 1994, an agreement was finally reached between the two countries over quotas and regulations of the industry. Also in 1994, Britain's Export Credit Guarantee Department began underwriting loans to Argentina again. The refusal of Argentina to purchase cattle products because of an outbreak in the UK of mad cow disease in the 1980s created new tensions between the two countries. Diplomatic relations have continued to improve with state visits to Argentina by foreign secretary Douglas Hurd and a year later a visit to Britain by foreign minister Guido di Tella of Argentina. In 1994 Britain's Duke of York was the first official

royal visitor to Argentina. Militarily, both countries joined forces to support the first Iraq war. Argentina and Great Britain are cooperating on the building and use of two nuclear power stations.

D324.M.250.E.E

Thirteen years after the war between Britain and Argentina over the Falkland Islands, Argentina still makes a ritual reference to Argentina's sovereignty over those islands. However, Britain continues to maintain sovereignty. In most other areas, relations between the two countries have steadily improved. Diplomatic relations were fully restored in 1985. In late 1992 and early the next year, major cabinet ministers visited each other countries. In 1994, the Duke of York, became the first royal visitor to Argentina since the Falklands War. Economic and commercial relations between the two countries steadily improved. Britain's Export Credits Guarantee Department resumed insurance cover to Argentina. That country hopes to attract UK investment in its nuclear industry and would have no objection to British companies operating its nuclear power stations. The two countries planned to develop oil resources in the South Atlantic jointly. A consortium led by British Gas bought the largest gas distribution company in Argentina. Although the UK and Argentina cooperated militarily during the Gulf War, Britain has maintained an arm embargo preventing Argentina from replacing aircraft shot down in the Falklands War. Argentina demanded an investigation of alleged war crimes during the Falklands War after reports of atrocities appeared in British media. Some differences and even anger exist between Argentina and Britain over territorial sea claims and fishing rights in the South Atlantic.

D324.M.250.E.F

Argentina was still obsessed with the Falkland Islands even in 1994, 12 years after its defeat in the 74-day war with Britain. The country's overriding foreign policy aim continued to be winning sovereignty over the islands. Relations between Argentina and Britain began improving after President Carlos Menem took office in 1989 and adopted pro-western foreign policies and free market economics. In theory, stronger trade and investment links with Argentina would gradually reduce the importance of the Falklands to Britain. Diplomatic relations resumed in 1990, and trade and investment between the two increased substantially. But the UK foreign office insisted its policy would change only if the islands' 2,000 inhabitants agreed. Argentina pressed for greater involvement in developing Falklands' natural resources, and in 1991, the UK agreed to cooperate to conserve fisheries. But Argentina began to issue fishing

licenses, muslcing in on the main source of revenue for the islands. Britain had hoped to strike an agreement with Argentina that would protect the islands' revenues, but Argentina tied such an agreement to concessions by London over shared development of the islands' possible oil reserves and lifting of the UK's arms embargo. British Gas and YPF, Argentina's state-owned oil company, began negotiating jointly exploring offshore gas and oilfields bordering the Falkland Islands in April, 1993. In January, 1994, the British indicated readiness to allow Argentine companies to take part in the development of oilfields in the Falkland Islands' territorial waters, provided they acknowledged that any oil extracted belonged to Britain.

D324.M.250.E.G

Argentine-British relations since the Falkland Islands War in 1982 have gradually improved. Argentine, however, steadfastly refuses to relinquish its claim to sovereignty over the Falklands. GB continues to refuse to discuss sovereignty formulas. This remains the only issue that seriously divides the two nations. Relations began to improve significantly when the newly elected president of Argentina, Carlos Menem, began to policy of rapprochement in 1989. His aim was to draw GB into a close relationship by opening talks on trade, oil and fishing rights, and military affairs. He began in August 1989 by lifting financial and trade restrictions on imports from Britain imposed during the Falkland Islands War. In early 1990, both countries agreed to restore full diplomatic ties and GB lifted a 150-mile military protection zone enforced around the islands. These developments were followed by a visit to London by Guido di Tella, Argentine foreign minister, to hold financial and trade discussions. In February 1992, both countries began discussions in Buenos Aires on developing South Atlantic oil resources. In December 1992, a British consortium bought out the Buenos Aires gas distribution company. Throughout 1993, a series of British high-level officials visited Argentina; notably, UK foreign secretary Douglas Hurd, the UK trade and industry secretary, and UK agriculture minister, Gillian Shephard, Argentina's economy minister, Domingo Cavallo, and its foreign minister, Guido di Tella, visited London. These exchanges resulted in greatly improved financial, oil, fisheries, and military issues but GB consistently refused to address the issue of sovereignty over the Falklands.

D324.M.250.E.H

Britain resumed trade with Argentina in 1985. In 1989 Argentina welcomed British imports. In 1990 diplomatic ties resumed. Britain lifted military protection zones around the islands. Each side gives advance notice of military exercises. Argentines

visit war cemeteries in the Falklands. Britain refutes Argentina's claims of sovereignty over the Falklands, disputed since 1833. Argentina's main objective remains recovering sovereignty of the Falklands, thinking economic links would reduce the Falkland's importance to Britain. Argentina "respects" the islands' "history", but won't recognize its local government. Britain's arms embargo continues, except for Argentine units participating with Gulf War Coalition forces. Argentine military officers train in the UK. Britain won't grant Argentine President Menem's request to visit the UK. Ministers have been exchanged. After 1990, rapid UK trade growth helped Argentina's "miracle" economic recovery. Argentina wants UK regulatory expertise in privatization and private sector investment in Argentina's gas and nuclear industries. British Gas bought the largest Argentine gas company and exploits off-shore gas with Argentine companies. A 1993 fisheries conservation agreement changed Argentine policies which threatened depletion of fragile fish populations. Britain angered Argentina by extending territorial waters to 200-miles around South Georgia and South Sandwich islands, which Argentina claims, and when it extended fisheries control within the Falkland's 200-mile limit. Argentina briefly banned UK cattle imports and opened investigations into allegations of Falkland War atrocities by British soldiers. In 1993 British business insurance covered UK companies in Argentina. In 1994 Britain indicated readiness for Argentine companies to participate in Falkland's off-shore oil development.

D324.M.250.E.I

In 1985 Britain ended trade curbs imposed on Argentina after the 1982 Falklands conflict. Argentina lifted restrictions on British imports in 1989. Full diplomatic relations were re-established in 1990, after Argentine president Carlos Menem took office. Reciprocal ministerial visits followed. Prince Andrew visited in 1994. Commercial relations improved steadily. Argentina sought UK expertise on privitization and agriculture. British export insurance resumed. British Gas bought into and managed Argentine gas distribution. UK's Babcock International supplied inspection equipment to an Argentine nuclear power station. Argentina used fishing disputes to press for oil and military compromises. Its cut-rate fishing licenses for the illex squid threatened Falkland income and overfishing. Britain extended territorial waters into disputed areas. A squid quota agreement was reached. Oil and gasfields detected in disputed waters required Argentine cooperation for development, but Britain claimed ownership of any oil. A British-imposed military protection zone around the Falklands was removed. Each side would announce upcoming military exercises. Britain and Argentina

co-operated during the 1991 Gulf War and Argentine units received UK-manufactured spare parts. This stopped. Britain blocked an Argentine military transport from landing in London, and dissuaded the US from selling fighter bombers to Argentina. Argentine officer training at British academies resumed. Argentinians visited war cemeteries in the Falklands and Scotland Yard investigated British war crimes in Argentina. Sovereignty was not discussed. Argentina did not recognize local Falklands government. Menem claimed rapprochement with Britain rather than confrontation would restore Argentina's Falklands sovereignty by 2000. London insisted co-operation did not imply sovereignty recognition.

D324.M.250.E.J

The return to normal relations between the United Kingdom and Argentina has been gradual since the end of the 1982 Falklands War. Full diplomatic ties were resumed in February 1990. In 1993, several high level visits were exchanged. Foreign Secretary Douglas Hurd and Agricultural Minister Gillian Shephard visited Argentina. Economy Minister Domingo Cavallo and Foreign Minister Guido De Tella went to the UK. In 1994, Britian's Prince Andrew made an official royal visit and Carlos Bastos, Argentine energy minister, went to London. Economic relations resumed much earlier and have been stronger. Britain ended trade curbs in 1985 and Argentina lifted restrictions in 1989. By 1991, British Gas was involved in gas production and distribution. Joint exploration was being discussed. Britain's Export Credit Guarantee Department resumed insurance for exports to Argentina. Other economic areas included livestock techniques, privatization of Argentina's nuclear industry, and sharing of fishing resources. Military cooperation was much slower. Britain did lift the 150-mile military protection zone around the Falklands in 1990, but the UK arms embargo, begun in 1982, continued through 1994. Talks over resuming training Argentine officers in UK military academies were planned The two nations agree to disagree over the sovereignty of the Falklands. Argentina's overriding aim is to regain sovereignty. British conservatives want no concessions and the policy is that the status will change only when the Falklanders, who do not trust Argentina, want a change. Key economic issues are selling licenses for the valuable illex squid fishing rights and oil exploration.

Appendix B

Pyramid Annotation for Document Collection D324

This appendix presents the Pyramid annotation of the Document Collection D324. Each SCU has a unique identifier (uid) and a label that represents the semantic content of the SCU. An SCU contains one or more contributors, which in turn contains one or more parts. The start and end attributes are character offsets into the set of human reference summaries.

<scu uid="25" label="As of October 1994, nothing had been resolved">
<contributor label="As of October 1994, nothing had been resolved">
<part label="As of October 1994, nothing had been resolved" start="1566"
end="1611"/>
</contributor>
<contributor label="Criminal investigations in Germany and the US relating to suspected industrial espionage, theft, perjury and wire fraud are still in progress">
<part label="Criminal investigations in Germany and the US relating to suspected industrial espionage, theft, perjury and wire fraud are still in progress" start="3044" end="3185"/>
</contributor>
<contributor label="Investigations continued...Through October, 1994, no legal action had been taken against Lopez or Volkswagen">
<part label="Through October, 1994, no legal action had been taken against Lopez or Volkswagen" start="4657" end="4738"/>
<part label="Investigations continued" start="4541" end="4565"/>
</contributor>

<contributor label="The outcome of the investigation is still uncertain">
<part label="The outcome of the investigation is still uncertain" start="6353" end="6404"/>
</contributor>
<contributor label="still unresolved investigations">
<part label="still unresolved investigations" start="7235" end="7266"/>
</contributor>
<contributor label="In October 1994 criminal charges were bogged down">
<part label="In October 1994 criminal charges were bogged down" start="9700" end="9749"/>
</contributor>
<contributor label="in October 1994 and no information was expected until a decision to indict Lopez was reached">
<part label="in October 1994 and no information was expected until a decision to indict Lopez was reached" start="11138" end="11230"/>
</contributor>
</scu>
<scu uid="5" label="Lopez left GM">
<contributor label="an employee of GM">
<part label="an employee of GM" start="155" end="172"/>
</contributor>
<contributor label="he left GM">
<part label="he left GM" start="1871" end="1881"/>
</contributor>
<contributor label="He left GM">
<part label="He left GM" start="3799" end="3809"/>
</contributor>
<contributor label="GM’s">
<part label="GM’s" start="5531" end="5535"/>
</contributor>
<contributor label="left his job...at General Motor’s">
<part label="left his job" start="6596" end="6608"/>
<part label="at General Motor’s" start="6631" end="6649"/>
</contributor>
<contributor label="he abruptly left GM">

<part label="he abruptly left GM" start="8582" end="8601"/>
</contributor>
<contributor label="GM">
<part label="GM" start="10015" end="10017"/>
</contributor>
</scu>
<scu uid="19" label="lopez took documents from GM to VW">
<contributor label="The situation became more serious when top- secret documents were found missing from GM">
<part label="The situation became more serious when top- secret documents were found missing from GM" start="528" end="615"/>
</contributor>
<contributor label="and allegedly took some sensitive GM documents and plans with him">
<part label="and allegedly took some sensitive GM documents and plans with him" start="2322" end="2387"/>
</contributor>
<contributor label="He also took, according to considerable evidence, many GM documents">
<part label="He also took, according to considerable evidence, many GM documents" start="3833" end="3900"/>
</contributor>
<contributor label="Mr. Lopez and his associates took GM and Adam Opel industrial secrets with them">
<part label="Mr. Lopez and his associates took GM and Adam Opel industrial secrets with them" start="5958" end="6037"/>
</contributor>
<contributor label="with secret GM documents he had requested">
<part label="with secret GM documents he had requested" start="8610" end="8651"/>
</contributor>
<contributor label="GM immediately accused Lopez of looting Opel’s supply network and contract database...GM documents and computerized information">
<part label="GM immediately accused Lopez of looting Opel’s supply network and contract database" start="10052" end="10135"/>

<part label="GM documents and computerized information" start="10517" end="10558"/>
</contributor>
<contributor label="Lopez also requested documents from Adam Opel that later turned up in the Wiesbaden home">
<part label="Lopez also requested documents from Adam Opel that later turned up in the Wiesbaden home" start="2389" end="2477"/>
</contributor>
</scu>
<scu uid="7" label="there was an industrial espionage case involving GM and VW">
<contributor label="The industrial espionage case involving GM and VW began with">
<part label="The industrial espionage case involving GM and VW began with" start="60" end="120"/>
</contributor>
<contributor label=""potentially the biggest-ever case of industrial espionage"">
<part label=""potentially the biggest-ever case of industrial espionage"" start="1979" end="2038"/>
</contributor>
<contributor label="The industrial espionage case involving Volkswagen and General Motors began">
<part label="The industrial espionage case involving Volkswagen and General Motors began" start="3243" end="3318"/>
</contributor>
<contributor label="The industrial espionage battle by General Motors (GM) and its German subsidiary, Adam Opel, against Volkswagen (VW)">
<part label="The industrial espionage battle by General Motors (GM) and its German subsidiary, Adam Opel, against Volkswagen (VW)" start="4796" end="4912"/>
</contributor>
<contributor label="into charges of industrial espionage">
<part label="into charges of industrial espionage" start="7267" end="7303"/>
</contributor>
<contributor label="investigations into industrial espionage">
<part label="investigations into industrial espionage" start="9291" end="9331"/>
</contributor>

<contributor label="and industrial espionage">
<part label="and industrial espionage" start="10337" end="10361"/>
</contributor>
</scu>
<scu uid="4" label="VW hired Jose Ignacio Lopez">
<contributor label="the hiring of Jose Ignacio Lopez...by VW">
<part label="the hiring of Jose Ignacio Lopez" start="121" end="153"/>
<part label="by VW" start="195" end="200"/>
</contributor>
<contributor label="for VW">
<part label="for VW" start="1882" end="1888"/>
</contributor>
<contributor label="for VW">
<part label="for VW" start="3810" end="3816"/>
</contributor>
<contributor label="The issue stems from...the alleged recruitment of...Jose Ignacio Lopez de Arriortura">
<part label="the alleged recruitment of" start="5504" end="5530"/>
<part label="Jose Ignacio Lopez de Arriortura" start="5590" end="5622"/>
<part label="The issue stems from" start="5483" end="5503"/>
</contributor>
<contributor label="Agnacio Lopes De Arriortua...to become Volkswagen’s">
<part label="to become Volkswagen’s" start="6664" end="6686"/>
<part label="Agnacio Lopes De Arriortua" start="6568" end="6594"/>
</contributor>
<contributor label="for VW">
<part label="for VW" start="8602" end="8608"/>
</contributor>
<contributor label="Lopez...moved to VW">
<part label="Lopez" start="9993" end="9998"/>
<part label="moved to VW" start="10029" end="10040"/>
</contributor>
</scu>
<scu uid="15" label="Ferdinand Piech is VW chairman">
<contributor label="VW Chairman Ferdinand Piech">

<part label="VW Chairman Ferdinand Piech" start="7676" end="7703"/>
</contributor>
<contributor label="Ferdinand Piech became Volkswagen chairman">
<part label="Ferdinand Piech became Volkswagen chairman" start="8201"
end="8243"/>
</contributor>
<contributor label="VW chairman Ferdinand Piech">
<part label="VW chairman Ferdinand Piech" start="9880" end="9907"/>
</contributor>
<contributor label="VW chairman, Ferdinand Piech">
<part label="VW chairman, Ferdinand Piech" start="5696" end="5724"/>
</contributor>
<contributor label="Ferdinand Piech, Chairman of Volkeswagen">
<part label="Ferdinand Piech, Chairman of Volkeswagen" start="3538" end="3578"/>
</contributor>
<contributor label="Ferdinand Piech had just been installed as Chairman of Volkswagen">
<part label="Ferdinand Piech had just been installed as Chairman of Volkswagen" start="225" end="290"/>
</contributor>
</scu>
<scu uid="6" label="seven other GM executives left with Lopez">
<contributor label="and seven other GM executives">
<part label="and seven other GM executives" start="9999" end="10028"/>
</contributor>
<contributor label="Lopez was followed by seven top members of his team">
<part label="Lopez was followed by seven top members of his team" start="8697" end="8748"/>
</contributor>
<contributor label="and seven of Lopez's business colleagues">
<part label="and seven of Lopez's business colleagues" start="5623" end="5663"/>
</contributor>
<contributor label="Lopez brought with him to VW seven former GM employees">
<part label="Lopez brought with him to VW seven former GM employees" start="312" end="366"/>

```
</contributor>
<contributor label="along with seven GM executives">
<part label="along with seven GM executives" start="3902" end="3932"/>
</contributor>
<contributor label="along with several of Lopez' key associates">
<part label="along with several of Lopez' key associates" start="2153" end="2196"/>
</contributor>
</scu>
<scu uid="29" label="The FBI is probing possible mail and wire fraud">
<contributor label="The FBI is probing possible mail and wire fraud">
<part label="The FBI is probing possible mail and wire fraud" start="5434"
end="5481"/>
</contributor>
<contributor label="and the FBI investigating mail and wire fraud">
<part label="and the FBI investigating mail and wire fraud" start="1374" end="1419"/>
</contributor>
<contributor label="Criminal investigations in...the US...wire fraud">
<part label="Criminal investigations in" start="3044" end="3070"/>
<part label="the US" start="3083" end="3089"/>
<part label="wire fraud" start="3153" end="3163"/>
</contributor>
<contributor label="The FBI opened still unresolved investigations of wire and mail
fraud against VW and Lopez">
<part label="The FBI opened still unresolved investigations of wire and mail fraud
against VW and Lopez" start="7498" end="7588"/>
</contributor>
<contributor label="and the FBI...and wire fraud">
<part label="and the FBI" start="9273" end="9284"/>
<part label="and wire fraud" start="9348" end="9362"/>
</contributor>
<contributor label="and the FBI began an investigation of mail and wire fraud">
<part label="and the FBI began an investigation of mail and wire fraud" start="11311"
end="11368"/>
</contributor>
</scu>
```

<scu uid="20" label="the missing documents described plans to build a new model car">
<contributor label="that described plans to build a new model car...including secrets documents containing plans to build the car">
<part label="that described plans to build a new model car" start="616" end="661"/>
<part label="including secrets documents containing plans to build the car" start="729" end="790"/>
</contributor>
<contributor label="investigators found details of Opel secret car plans">
<part label="investigators found details of Opel secret car plans" start="4104" end="4156"/>
</contributor>
<contributor label="and information on new models">
<part label="and information on new models" start="6193" end="6222"/>
</contributor>
<contributor label="designs for advanced cars...and engines;">
<part label="designs for advanced cars" start="6820" end="6845"/>
<part label="and engines;" start="6846" end="6858"/>
</contributor>
<contributor label="and car models">
<part label="and car models" start="8681" end="8695"/>
</contributor>
<contributor label="and a new Opel mini-car">
<part label="and a new Opel mini-car" start="10185" end="10208"/>
</contributor>
</scu>
<scu uid="21" label="documents included secret plans for a new factory">
<contributor label="that detailed GM’s new plant">
<part label="that detailed GM’s new plant" start="8652" end="8680"/>
</contributor>
<contributor label="and seized documents that were later founds to contain Opel plans">
<part label="and seized documents that were later founds to contain Opel plans" start="2572" end="2637"/>
</contributor>
<contributor label="plans for a new style...car factory">

<part label="plans for a new style" start="6137" end="6158"/>
<part label="car factory" start="6180" end="6191"/>
</contributor>
<contributor label="factories;">
<part label="factories;" start="6809" end="6819"/>
</contributor>
<contributor label="and taking secret plans for a...factory">
<part label="and taking secret plans for a" start="10136" end="10165"/>
<part label="factory" start="10177" end="10184"/>
</contributor>
</scu>
<scu uid="14" label="Ferdinand Piech recruited the General Motors/Opel executive, Jose Lopez de Arriortua">
<contributor label="Ferdinand Piech recruited the General Motors/Opel executive, Jose Lopez de Arriortua">
<part label="Ferdinand Piech recruited the General Motors/Opel executive, Jose Lopez de Arriortua" start="9892" end="9976"/>
</contributor>
<contributor label="he recruited">
<part label="he recruited" start="8338" end="8350"/>
</contributor>
<contributor label="when he hired Lopez">
<part label="when he hired Lopez" start="291" end="310"/>
</contributor>
<contributor label="He is accused of luring Lopez away from GM">
<part label="He is accused of luring Lopez away from GM" start="2110" end="2152"/>
</contributor>
<contributor label="He was presumably recruited by VW chairman, Ferdinand Piech">
<part label="He was presumably recruited by VW chairman, Ferdinand Piech" start="5665" end="5724"/>
</contributor>
</scu>
<scu uid="30" label="Lopez paid a fine instead of facing perjury charges">
<contributor label="Lopez paid a fine instead of facing perjury charges">

<part label="Lopez paid a fine instead of facing perjury charges" start="9582" end="9633"/>
</contributor>
<contributor label="to pay a DM75,000 fine to avoid facing charges in court">
<part label="to pay a DM75,000 fine to avoid facing charges in court" start="10922" end="10977"/>
</contributor>
<contributor label="So far, Lopez has agreed to pay DM75,000 to set aside the perjury case against him">
<part label="So far, Lopez has agreed to pay DM75,000 to set aside the perjury case against him" start="8061" end="8143"/>
</contributor>
<contributor label="Lopez agreed to pay 29,850 pounds to avoid facing perjury charges in court">
<part label="Lopez agreed to pay 29,850 pounds to avoid facing perjury charges in court" start="4581" end="4655"/>
</contributor>
<contributor label="Lopez agreed to pay DM75,000 instead of facing perjury charges in court">
<part label="Lopez agreed to pay DM75,000 instead of facing perjury charges in court" start="2971" end="3042"/>
</contributor>
</scu>
<scu uid="26" label="Lopez was accused of perjury">
<contributor label="Lopez was accused of perjury">
<part label="Lopez was accused of perjury" start="10836" end="10864"/>
</contributor>
<contributor label="perjury">
<part label="perjury" start="9340" end="9347"/>
</contributor>
<contributor label="and perjury against Lopez">
<part label="and perjury against Lopez" start="7312" end="7337"/>
</contributor>
<contributor label="instead of facing perjury charges in court">
<part label="instead of facing perjury charges in court" start="3000" end="3042"/>

</contributor>
<contributor label="to avoid facing perjury charges in court">
<part label="to avoid facing perjury charges in court" start="4615" end="4655"/>
</contributor>
</scu>
<scu uid="11" label="Lopez was GMs' procurement chief">
<contributor label="procurement chief">
<part label="procurement chief" start="5572" end="5589"/>
</contributor>
<contributor label="General Motors' likeminded global head of purchasing Jose Ignacio Lopez de Arriortua">
<part label="General Motors' likeminded global head of purchasing Jose Ignacio Lopez de Arriortua" start="8351" end="8435"/>
</contributor>
<contributor label="as head of purchasing">
<part label="as head of purchasing" start="6609" end="6630"/>
</contributor>
<contributor label="Jose Lopez as head of purchasing...He was made procurement chief at GM headquarters">
<part label="Jose Lopez as head of purchasing" start="1669" end="1701"/>
<part label="He was made procurement chief at GM headquarters" start="1812" end="1860"/>
</contributor>
<contributor label="Lopez was procurement chief at Adam Opel">
<part label="Lopez was procurement chief at Adam Opel" start="3609" end="3649"/>
</contributor>
</scu>
<scu uid="12" label=" Adam Opel is subsidiary of GM">
<contributor label="subsidiary Adam Opel">
<part label="subsidiary Adam Opel" start="173" end="193"/>
</contributor>
<contributor label="at Adam Opel, GMs German subsidiary">
<part label="at Adam Opel, GMs German subsidiary" start="1702" end="1737"/>
</contributor>
<contributor label="Adam Opel, GM's German subsidiary">

<part label="Adam Opel, GM's German subsidiary" start="3640" end="3673"/>
</contributor>
<contributor label="General Motor's Opel,Germany">
<part label="General Motor's Opel,Germany" start="6634" end="6662"/>
</contributor>
</scu>
<scu uid="23" label="documents were found where former GM employee were staying">
<contributor label="later turned up in the Wiesbaden home of a Lopez colleague who followed him to VW">
<part label="later turned up in the Wiesbaden home of a Lopez colleague who followed him to VW" start="2440" end="2521"/>
</contributor>
<contributor label="and documents were found at the apartment of the former GM executives">
<part label="and documents were found at the apartment of the former GM executives" start="10594" end="10663"/>
</contributor>
<contributor label="two Lopez associates">
<part label="two Lopez associates" start="4057" end="4077"/>
</contributor>
<contributor label="another former GM employee were staying">
<part label="another former GM employee were staying" start="819" end="858"/>
</contributor>
</scu>
<scu uid="24" label="German officials began investigating VW for theft">
<contributor label="German officials began investigating VW for theft">
<part label="German officials began investigating VW for theft" start="10287" end="10336"/>
</contributor>
<contributor label="German state prosecutors...began investigations into...theft">
<part label="German state prosecutors" start="9248" end="9272"/>
<part label="began investigations into" start="9285" end="9310"/>
<part label="theft" start="9333" end="9338"/>
</contributor>

<contributor label="Germany investigator, Dorothea Holland, launched...investigations into charges of...theft">
<part label="Germany investigator, Dorothea Holland, launched" start="7186" end="7234"/>
<part label="investigations into charges of" start="7252" end="7282"/>
<part label="theft" start="7305" end="7310"/>
</contributor>
<contributor label="The investigation is focused mainly on evidence that">
<part label="The investigation is focused mainly on evidence that" start="5905" end="5957"/>
</contributor>
</scu>
<scu uid="43" label="Lopez left Opel On March 16, 1993">
<contributor label="On March 16, 1993">
<part label="On March 16, 1993" start="6462" end="6479"/>
</contributor>
<contributor label="but when he learned in March 1993">
<part label="but when he learned in March 1993" start="8514" end="8547"/>
</contributor>
<contributor label="In March 1993...overnight">
<part label="In March 1993" start="9978" end="9991"/>
<part label="overnight" start="10041" end="10050"/>
</contributor>
<contributor label="in March, 1993">
<part label="in March, 1993" start="3817" end="3831"/>
</contributor>
</scu>
<scu uid="10" label="Lopez was Basque-born">
<contributor label="Basque-born...in his Basque area">
<part label="Basque-born" start="5560" end="5571"/>
<part label="in his Basque area" start="6319" end="6337"/>
</contributor>
<contributor label="in his native Basque country">
<part label="in his native Basque country" start="2257" end="2285"/>
</contributor>

<contributor label="in his own Basque country">
<part label="in his own Basque country" start="3748" end="3773"/>
</contributor>
<contributor label="in his native Basque country">
<part label="in his native Basque country" start="8484" end="8512"/>
</contributor>
</scu>
<scu uid="52" label="Lopez was disappointed by GM's decision not to build an automobile plant in his own Basque country">
<contributor label="Lopez was disappointed by GM's decision not to build an automobile plant in his own Basque country">
<part label="Lopez was disappointed by GM's decision not to build an automobile plant in his own Basque country" start="3675" end="3773"/>
</contributor>
<contributor label="Lopez, disappointed that GM was not going to build a plant in his native Basque country">
<part label="Lopez, disappointed that GM was not going to build a plant in his native Basque country" start="2198" end="2285"/>
</contributor>
<contributor label="Coincidentally, Lopez quit after being informed that a plan to install his new car dream plant in his Basque area was cancelled">
<part label="Coincidentally, Lopez quit after being informed that a plan to install his new car dream plant in his Basque area was cancelled" start="6224" end="6351"/>
</contributor>
<contributor label="Lopez had developed a new GM plant to be built in his native Basque country...that it would be built in Hungary">
<part label="Lopez had developed a new GM plant to be built in his native Basque country" start="8437" end="8512"/>
<part label="that it would be built in Hungary" start="8548" end="8581"/>
</contributor>
</scu>
<scu uid="8" label="VW and Lopez also were accused on conducting an illegal recruiting campaign">
<contributor label="VW and Lopez also were accused on conducting an illegal recruiting campaign">

<part label="VW and Lopez also were accused on conducting an illegal recruiting campaign" start="10210" end="10285"/>
</contributor>
<contributor label="personnel poaching">
<part label="personnel poaching" start="9490" end="9508"/>
</contributor>
<contributor label="charges of anti-competitive staff poaching">
<part label="charges of anti-competitive staff poaching" start="7009" end="7051"/>
</contributor>
<contributor label="the recruiting of their employees">
<part label="the recruiting of their employees" start="437" end="470"/>
</contributor>
</scu>
<scu uid="287" label=" Gutierrez and Piazza were the former GM associates who were found with the plans">
<contributor label="where Gutierrez and Piazza">
<part label="where Gutierrez and Piazza" start="791" end="817"/>
</contributor>
<contributor label="Jorge Alvarez Aquirre and Rosario Piazza">
<part label="Jorge Alvarez Aquirre and Rosario Piazza" start="4015" end="4055"/>
</contributor>
<contributor label="Jorge Alvarez Aquirre and Rosario Piazza">
<part label="Jorge Alvarez Aquirre and Rosario Piazza" start="10665" end="10705"/>
</contributor>
</scu>
<scu uid="40" label="A regional court in Frankfurt issued an injunction preventing VW from recruiting more GM staff">
<contributor label="A regional court in Frankfurt issued an injunction preventing VW from recruiting more GM staff">
<part label="A regional court in Frankfurt issued an injunction preventing VW from recruiting more GM staff" start="6909" end="7003"/>
</contributor>
<contributor label="With GM urging, a temporary injunction was imposed on VW recruiting">

<part label="With GM urging, a temporary injunction was imposed on VW recruiting" start="10363" end="10430"/>
</contributor>
<contributor label="VW was banned from further personnel poaching but a Frankfurt court">
<part label="VW was banned from further personnel poaching but a Frankfurt court" start="9463" end="9530"/>
</contributor>
</scu>
<scu uid="41" label="All charges of anti-competitive staff poaching were later dismissed">
<contributor label="All charges of anti-competitive staff poaching were later dismissed">
<part label="All charges of anti-competitive staff poaching were later dismissed" start="7005" end="7072"/>
</contributor>
<contributor label="but it was subsequently lifted and manager-poaching claims against VW were rejected">
<part label="but it was subsequently lifted and manager-poaching claims against VW were rejected" start="10432" end="10515"/>
</contributor>
<contributor label="but a Frankfurt court denied that poaching broke fair competition rules">
<part label="but a Frankfurt court denied that poaching broke fair competition rules" start="9509" end="9580"/>
</contributor>
</scu>
<scu uid="44" label="Ferdinand Piech took over an ailing VW company that was losing money">
<contributor label="Ferdinand Piech took over an ailing VW company that was losing money">
<part label="Ferdinand Piech took over an ailing VW company that was losing money" start="2040" end="2108"/>
</contributor>

<contributor label="He saw Lopez as the answer to a similar ongoing problem at VW">
<part label="He saw Lopez as the answer to a similar ongoing problem at VW" start="5842" end="5903"/>
</contributor>
<contributor label="and planned to turn the money-losing company around">
<part label="and planned to turn the money-losing company around" start="8260" end="8311"/>
</contributor>
</scu>
<scu uid="35" label="The U.S. Justice Department’s interest in industrial espionage had been piqued">
<contributor label="The U.S. Justice Department’s interest in industrial espionage had been piqued">
<part label="The U.S. Justice Department’s interest in industrial espionage had been piqued" start="11232" end="11310"/>
</contributor>
<contributor label="In July, the U.S. Justice Department announced it was investigating the Lopez case">
<part label="In July, the U.S. Justice Department announced it was investigating the Lopez case" start="4225" end="4307"/>
</contributor>
<contributor label="including the US Justice Dept">
<part label="including the US Justice Dept" start="1343" end="1372"/>
</contributor>
</scu>
<scu uid="58" label=" Gunter Rexrodt is the German economics minister">
<contributor label="The German economics minister, Gunter Rexrodt">
<part label="The German economics minister, Gunter Rexrodt" start="10707" end="10752"/>
</contributor>
<contributor label="German economics minister Gunter Rexrodt">
<part label="German economics minister Gunter Rexrodt" start="2731" end="2771"/>
</contributor>
</scu>

<scu uid="27" label="Documents included details of Opel's entire European component supplier network and key contact data">
<contributor label="These included details of Opel's entire European component supplier network and key contact data">
<part label="These included details of Opel's entire European component supplier network and key contact data" start="6039" end="6135"/>
</contributor>
<contributor label="and information about Opel's suppliers and parts">
<part label="and information about Opel's suppliers and parts" start="6859" end="6907"/>
</contributor>
</scu>
<scu uid="90" label="Eurothere are fears of destabilization of relations between Germany and America">
<contributor label="European government and industry leaders expressed fear...would destabilize U.S.-European commercial and diplomatic relations">
<part label="European government and industry leaders expressed fear" start="7590" end="7645"/>
<part label="would destabilize U.S.-European commercial and diplomatic relations" start="7782" end="7849"/>
</contributor>
<contributor label="is now also worried about Bonn's relations with Washington">
<part label="is now also worried about Bonn's relations with Washington" start="5170" end="5228"/>
</contributor>
</scu>
<scu uid="82" label="German newspapers such as Der Speigel made public allegations of spying against Lopez">
<contributor label="for making public allegations of spying against Lopez">
<part label="for making public allegations of spying against Lopez" start="1139" end="1192"/>
</contributor>
<contributor label="and details were aired in German newspapers">
<part label="and details were aired in German newspapers" start="2686" end="2729"/>
</contributor>

</scu>
<scu uid="65" label="GM charged that during his last months at GM, Lopez stole GM plans">
<contributor label="GM charged that during his last months at GM, Lopez stole GM plans">
<part label="GM charged that during his last months at GM, Lopez stole GM plans" start="6723" end="6789"/>
</contributor>
<contributor label="At GM's request">
<part label="At GM's request" start="9231" end="9246"/>
</contributor>
</scu>
<scu uid="57" label="Gunter Rexrodt, had tried to be a peacemaker is this controversy">
<contributor label="Gunter Rexrodt, had tried to be a peacemaker is this controversy">
<part label="Gunter Rexrodt, had tried to be a peacemaker is this controversy" start="10738" end="10802"/>
</contributor>
<contributor label="and for a time tried to be a peacemaker">
<part label="and for a time tried to be a peacemaker" start="2842" end="2881"/>
</contributor>
</scu>
<scu uid="69" label="investigation is bogged down in political and legal translatlantic issues">
<contributor label="for the past 18 months has bogged down in mountains of paper and a complex transatlantic tussle involving both lawyers and politicians">
<part label="for the past 18 months has bogged down in mountains of paper and a complex transatlantic tussle involving both lawyers and politicians" start="4913" end="5047"/>
</contributor>
<contributor label="The legal cases soon became bogged down in mountains of papers and transatlantic issues between the countries involving lawyers and politicians">
<part label="The legal cases soon became bogged down in mountains of papers and transatlantic issues between the countries involving lawyers and politicians" start="1421" end="1564"/>

</contributor>
</scu>
<scu uid="28" label="Lopez helped turn around Opel">
<contributor label="led the company to become the most profitable car maker in that country">
<part label="led the company to become the most profitable car maker in that country" start="1739" end="1810"/>
</contributor>
<contributor label="Lopez's leading role in helping Adam Opel recover from a major production cost disadvantage">
<part label="Lopez's leading role in helping Adam Opel recover from a major production cost disadvantage" start="5749" end="5840"/>
</contributor>
</scu>
<scu uid="66" label="Lopez paid fine in May, 1994">
<contributor label="In May, 1994">
<part label="In May, 1994" start="4567" end="4579"/>
</contributor>
<contributor label="and in May 1994 agreed, while maintaining his innocence">
<part label="and in May 1994 agreed, while maintaining his innocence" start="10865" end="10920"/>
</contributor>
</scu>
<scu uid="16" label="Lopez was hired as VW production director">
<contributor label="as production director">
<part label="as production director" start="201" end="223"/>
</contributor>
<contributor label="to become Volkswagen's Purchasing and Production Director">
<part label="to become Volkswagen's Purchasing and Production Director" start="6664" end="6721"/>
</contributor>
</scu>
<scu uid="87" label="Piech got nationalistic in his accusations">
<contributor label="Piech's damaging nationalistic tones">
<part label="Piech's damaging nationalistic tones" start="9136" end="9172"/>

</contributor>
<contributor label="that the Lopez incident amounted to U.S. industrial warfare against Germany">
<part label="that the Lopez incident amounted to U.S. industrial warfare against Germany" start="7705" end="7780"/>
</contributor>
</scu>
<scu uid="53" label="Still Later German police raided VW headquarters">
<contributor label="Still Later German police raided VW headquarters">
<part label="Still Later German police raided VW headquarters" start="2523" end="2571"/>
</contributor>
<contributor label="were seized from a VW headquarters">
<part label="were seized from a VW headquarters" start="10559" end="10593"/>
</contributor>
</scu>
<scu uid="285" label="The factory in stolen plans was high-speed">
<contributor label="high-speed">
<part label="high-speed" start="10166" end="10176"/>
</contributor>
<contributor label="high-speed">
<part label="high-speed" start="6169" end="6179"/>
</contributor>
</scu>
<scu uid="286" label="The factory in stolen plans was low-cost">
<contributor label="low-cost">
<part label="low-cost" start="6160" end="6168"/>
</contributor>
<contributor label="for ultra-low cost">
<part label="for ultra-low cost" start="6790" end="6808"/>
</contributor>
</scu>
<scu uid="77" label="The German prosecutor was Dorthea Holland">
<contributor label="The German prosecutor, Dorthea Holland">
<part label="The German prosecutor, Dorthea Holland" start="10978" end="11016"/>

</contributor>
<contributor label="Darmstadt, Germany investigator, Dorothea Holland">
<part label="Darmstadt, Germany investigator, Dorothea Holland" start="7175" end="7224"/>
</contributor>
</scu>
<scu uid="22" label="Then state prosecution officials discovered four boxes of papers">
<contributor label="Then state prosecution officials discovered four boxes of papers">
<part label="Then state prosecution officials discovered four boxes of papers" start="663" end="727"/>
</contributor>
<contributor label="In four remaining boxes">
<part label="In four remaining boxes" start="4079" end="4102"/>
</contributor>
</scu>
<scu uid="32" label="There were investigations and counter charges on both sides of the ocean">
<contributor label="This led to investigations and counter charges on both sides of the ocean">
<part label="This led to investigations and counter charges on both sides of the ocean" start="860" end="933"/>
</contributor>
<contributor label="Months of charges and counter-charges followed">
<part label="Months of charges and counter-charges followed" start="2639" end="2685"/>
</contributor>
</scu>
<scu uid="70" label="VW counter-charged that GM had planted GM documents and data in VW sites and computers">
<contributor label="VW counter-charged that GM had planted GM documents and data in VW sites and computers">
<part label="VW counter-charged that GM had planted GM documents and data in VW sites and computers" start="7410" end="7496"/>
</contributor>

<contributor label="Piech publicly accused GM/Opel of planting documents and hacking VW computers with the aim of destroying VW">
<part label="Piech publicly accused GM/Opel of planting documents and hacking VW computers with the aim of destroying VW" start="8971" end="9078"/>
</contributor>
</scu>
<scu uid="59" label=" in September 1993, Rexrodt withdrew as peacemaker">
<contributor label="but in September 1993 withdrew">
<part label="but in September 1993 withdrew" start="10804" end="10834"/>
</contributor>
</scu>
<scu uid="50" label=" VW failed to convince GM that its plans were not plagiarized">
<contributor label="when VW failed to convince GM that its plans for a revolutionary automobile plant in Spain were not copies of a proposed GM project">
<part label="when VW failed to convince GM that its plans for a revolutionary automobile plant in Spain were not copies of a proposed GM project" start="3319" end="3450"/>
</contributor>
</scu>
<scu uid="295" label="A court case was also brought against leading news magazine, Der Spiegel">
<contributor label="A court case was also brought against leading news magazine, Der Spiegel">
<part label="A court case was also brought against leading news magazine, Der Spiegel" start="1065" end="1137"/>
</contributor>
</scu>
<scu uid="36" label="A US probe of the investigation started at the instigation of the Commerce Department">
<contributor label="A US probe of the investigation started at the instigation of the Commerce Department">
<part label="A US probe of the investigation started at the instigation of the Commerce Department" start="5230" end="5315"/>
</contributor>
</scu>

<scu uid="60" label="A VW employee said she had punched Opel data into the VW computer">
<contributor label="A VW employee said she had punched Opel data into the VW computer">
<part label="A VW employee said she had punched Opel data into the VW computer" start="4158" end="4223"/>
</contributor>
</scu>
<scu uid="51" label="As early as December, 1992, Lopez was in touch with Piech">
<contributor label="As early as December, 1992, Jose Ignacio Lopez de Arriortua, of GM, was in touch with Ferdinand Piech...about coming to work for VW">
<part label="As early as December, 1992, Jose Ignacio Lopez de Arriortua, of GM, was in touch with Ferdinand Piech" start="3452" end="3553"/>
<part label="about coming to work for VW" start="3580" end="3607"/>
</contributor>
</scu>
<scu uid="67" label="At a VW meeting in August, 1993, Lopez contradicted his earlier public claim that he never took any secret documents">
<contributor label="At a VW meeting in August, 1993, Lopez contradicted his earlier public claim that he never took any secret documents">
<part label="At a VW meeting in August, 1993, Lopez contradicted his earlier public claim that he never took any secret documents" start="4309" end="4425"/>
</contributor>
</scu>
<scu uid="78" label="Because of leaks, a gag was placed on Holland’s office">
<contributor label="Because of leaks, a gag was placed on her office">
<part label="Because of leaks, a gag was placed on her office" start="11089" end="11137"/>
</contributor>
</scu>
<scu uid="394" label="defensive allegations were made by VW Chairman Ferdinand Piech">
<contributor label="defensive allegations by VW Chairman Ferdinand Piech">
<part label="defensive allegations by VW Chairman Ferdinand Piech" start="7651" end="7703"/>

</contributor>
</scu>
<scu uid="296" label="Der Spiegal is a leading news magazine">
<contributor label="leading news magazine, Der Spiegel">
<part label="leading news magazine, Der Spiegel" start="1103" end="1137"/>
</contributor>
</scu>
<scu uid="83" label="Der Spiegel later presented evidence in state court in a bid not to be stopped from reporting">
<contributor label="Der Spiegel later presented evidence in state court in a bid not to be stopped from reporting">
<part label="Der Spiegel later presented evidence in state court in a bid not to be stopped from reporting" start="1194" end="1287"/>
</contributor>
</scu>
<scu uid="42" label="Documents were shredded">
<contributor label="In April, 1993, witnesses in Wiesbaden allegedly saw documents being shredded by">
<part label="In April, 1993, witnesses in Wiesbaden allegedly saw documents being shredded by" start="3934" end="4014"/>
</contributor>
</scu>
<scu uid="76" label="Dorthea Holland, was searching through an estimated 2 million computer printout sheets">
<contributor label="Dorthea Holland, was searching through an estimated 2 million computer printout sheets">
<part label="Dorthea Holland, was searching through an estimated 2 million computer printout sheets" start="11001" end="11087"/>
</contributor>
</scu>
<scu uid="34" label="FBI investigation was also stalled">
<contributor label="which was also stalled">
<part label="which was also stalled" start="11370" end="11392"/>
</contributor>
</scu>

<scu uid="56" label="General Motors Corporation and Volkswagen were warring in 1993 and 1994">
<contributor label="General Motors Corporation and Volkswagen were warring in 1993 and 1994">
<part label="General Motors Corporation and Volkswagen were warring in 1993 and 1994" start="9807" end="9878"/>
</contributor>
</scu>
<scu uid="392" label="Gerhardt Schroeder is from Lower Saxony">
<contributor label="Gerhardt Schroeder of Lower Saxony">
<part label="Gerhardt Schroeder of Lower Saxony" start="2886" end="2920"/>
</contributor>
</scu>
<scu uid="63" label="Gerhardt Schroeder supported VW">
<contributor label="PM Gerhardt Schroeder...strongly supported VW">
<part label="strongly supported VW" start="2948" end="2969"/>
<part label="PM Gerhardt Schroeder" start="2883" end="2904"/>
</contributor>
</scu>
<scu uid="288" label="German politicians called the case biased">
<contributor label="German politicians called the case biased">
<part label="German politicians called the case biased" start="9635" end="9676"/>
</contributor>
</scu>
<scu uid="289" label="Germans wanted the case dropped">
<contributor label="and wanted it dropped">
<part label="and wanted it dropped" start="9677" end="9698"/>
</contributor>
</scu>
<scu uid="298" label="Germany distanced itself from Piech’s accusations">
<contributor label="and led Germany to distance itself from">
<part label="and led Germany to distance itself from" start="9096" end="9135"/>
</contributor>
</scu>

<scu uid="94" label="Germany is concerned about the effect of the court investigations on domestic economic and political affairs">
<contributor label="Germany, increasingly concerned about the effect of the court investigations on domestic economic and political affairs">
<part label="Germany, increasingly concerned about the effect of the court investigations on domestic economic and political affairs" start="5049" end="5168"/>
</contributor>
</scu>
<scu uid="17" label="GM employees leaving for VW included Lopez's close friend, Jorge Manuel Gutierrez">
<contributor label="including his close friend, Jorge Manuel Gutierrez">
<part label="including his close friend, Jorge Manuel Gutierrez" start="367" end="417"/>
</contributor>
</scu>
<scu uid="55" label="GM settled for only some employess to be banned from working for VW">
<contributor label="but settled for only some">
<part label="but settled for only some" start="9436" end="9461"/>
</contributor>
</scu>
<scu uid="54" label="GM wanted all former employees banned from working for VW for 12 months">
<contributor label="GM wanted all former employees banned from working for VW for 12 months">
<part label="GM wanted all former employees banned from working for VW for 12 months" start="9364" end="9435"/>
</contributor>
</scu>
<scu uid="62" label="Gunter Rexrodt was concerned of damage to US-German relations">
<contributor label="Gunter Rexrodt was concerned of damage to US-German political and business relations">
<part label="Gunter Rexrodt was concerned of damage to US-German political and business relations" start="2757" end="2841"/>

</contributor>
</scu>
<scu uid="85" label="investigations in both countries were followed by civil and criminal court cases">
<contributor label="followed by civil and criminal court cases">
<part label="followed by civil and criminal court cases" start="934" end="976"/>
</contributor>
</scu>
<scu uid="292" label="investigations were launched against Lopez's 22-year old daughter">
<contributor label="his 22-year old daughter">
<part label="his 22-year old daughter" start="7339" end="7363"/>
</contributor>
</scu>
<scu uid="294" label="investigations were launched against other GM colleagues now at VW">
<contributor label="and other GM colleagues now at VW">
<part label="and other GM colleagues now at VW" start="7375" end="7408"/>
</contributor>
</scu>
<scu uid="38" label="Japanese car import quotas to Europe expire in two years">
<contributor label="with Japanese car import quotas to Europe expiring in two years">
<part label="with Japanese car import quotas to Europe expiring in two years" start="6481" end="6544"/>
</contributor>
</scu>
<scu uid="80" label="Lopez faced further charges">
<contributor label="but faced further charges">
<part label="but faced further charges" start="1038" end="1063"/>
</contributor>
</scu>
<scu uid="86" label="Lopez left GM under a cloud of confusion">
<contributor label="However...under circumstances, which along with ensuing events, were described by a German judge as...left GM under a cloud of confusion">
<part label="However" start="1862" end="1869"/>

<part label="under circumstances, which along with ensuing events, were described by a German judge as" start="1889" end="1978"/>
<part label="left GM under a cloud of confusion" start="2287" end="2321"/>
</contributor>
</scu>
<scu uid="68" label="Lopez said that papers from his former offices were destroyed in order to keep them from being circulated within VW">
<contributor label="and said that papers from his former offices were destroyed in order to keep them from being circulated within VW">
<part label="and said that papers from his former offices were destroyed in order to keep them from being circulated within VW" start="4426" end="4539"/>
</contributor>
</scu>
<scu uid="48" label="Lopez tried to recruit others">
<contributor label="He tried to recruit others">
<part label="He tried to recruit others" start="8819" end="8845"/>
</contributor>
</scu>
<scu uid="39" label="Lopez was a renowned cost-cutter">
<contributor label="renowned cost-cutter">
<part label="renowned cost-cutter" start="6546" end="6566"/>
</contributor>
</scu>
<scu uid="9" label="Lopez was eccentric and visionary">
<contributor label="eccentric and visionary">
<part label="eccentric and visionary" start="5536" end="5559"/>
</contributor>
</scu>
<scu uid="79" label="Lopez was found innocent during his first trial in Germany">
<contributor label="Lopez was found innocent during his first trial in Germany">
<part label="Lopez was found innocent during his first trial in Germany" start="978" end="1036"/>
</contributor>
</scu>
<scu uid="293" label="Lopez's daughter Begounia is 22 years old">

<contributor label="Begounia">
<part label="Begounia" start="7365" end="7373"/>
</contributor>
</scu>
<scu uid="393" label="Lower Saxony is VW's largest shareholder">
<contributor label="VW's largest shareholder">
<part label="VW's largest shareholder" start="2922" end="2946"/>
</contributor>
</scu>
<scu uid="47" label="people leaving GM with Lopez included Jose Gutierrez, Jorge Alvarez Aguirre, and Rosario Piazza">
<contributor label="including Jose Gutierrez, Jorge Alvarez Aguirre, and Rosario Piazza">
<part label="including Jose Gutierrez, Jorge Alvarez Aguirre, and Rosario Piazza" start="8750" end="8817"/>
</contributor>
</scu>
<scu uid="45" label="Piech became VW chairman in January 1993">
<contributor label="in January 1993">
<part label="in January 1993" start="8244" end="8259"/>
</contributor>
</scu>
<scu uid="49" label="Piech was a ruthless restructurer">
<contributor label="A ruthless restructurer">
<part label="A ruthless restructurer" start="8313" end="8336"/>
</contributor>
</scu>
<scu uid="75" label="Piech was impressed with Lopez">
<contributor label="who was impressed with">
<part label="who was impressed with" start="5726" end="5748"/>
</contributor>
</scu>
<scu uid="72" label="Piech's clumsy, halfhearted conciliation efforts failed">
<contributor label="Piech's clumsy, halfhearted conciliation efforts failed">

<part label="Piech's clumsy, halfhearted conciliation efforts failed" start="9174" end="9229"/>
</contributor>
</scu>
<scu uid="297" label="Piecs's accusations angered GM">
<contributor label="This angered GM">
<part label="This angered GM" start="9080" end="9095"/>
</contributor>
</scu>
<scu uid="37" label="President Clinton decided that industrial espionage was a threat to America's well being">
<contributor label="after President Clinton apparently decided that industrial espionage in general was a threat to America's well being">
<part label="after President Clinton apparently decided that industrial espionage in general was a threat to America's well being" start="5316" end="5432"/>
</contributor>
</scu>
<scu uid="92" label="Ron Brown, suggested that relations between the U.S. and Germany would be damaged">
<contributor label="Ron Brown, suggested that relations between the U.S. and Germany would be damaged">
<part label="Ron Brown, suggested that relations between the U.S. and Germany would be damaged" start="7876" end="7957"/>
</contributor>
</scu>
<scu uid="291" label="Soon after Lopez's arrival, GM planned a car similar to a planned GM model">
<contributor label="and a car similar to a planned GM model">
<part label="and a car similar to a planned GM model" start="8930" end="8969"/>
</contributor>
</scu>
<scu uid="290" label="Soon after Lopez's arrival, VW announced a new plant to be built in Basque country">
<contributor label="Soon after Lopez's arrival, VW announced a new plant to be built in Basque country">

<part label="Soon after Lopez's arrival, VW announced a new plant to be built in Basque country" start="8847" end="8929"/>
</contributor>
</scu>
<scu uid="84" label="The case reached the highest levels in both countries">
<contributor label="The case reached the highest levels in both countries">
<part label="The case reached the highest levels in both countries" start="1289" end="1342"/>
</contributor>
</scu>
<scu uid="91" label="U.S. Commerce Secretary is Ron Brown">
<contributor label="U.S. Commerce Secretary, Ron Brown">
<part label="U.S. Commerce Secretary, Ron Brown" start="7851" end="7885"/>
</contributor>
</scu>
<scu uid="93" label="US wants German investigatorsto immediately deliver long promised data and assistance in the GM/VW case">
<contributor label="if German investigators don't immediately deliver long promised data and assistance in the GM/VW case">
<part label="if German investigators don't immediately deliver long promised data and assistance in the GM/VW case" start="7958" end="8059"/>
</contributor>
</scu>
<scu uid="73" label="VW failed to get court injunctions preventing Der Spiegel magazine from publishing GM's allegations">
<contributor label="VW failed to get court injunctions preventing Der Spiegel magazine from publishing GM's allegations">
<part label="VW failed to get court injunctions preventing Der Spiegel magazine from publishing GM's allegations" start="7074" end="7173"/>
</contributor>
</scu>
<scu uid="61" label="VW offered to build the Basque plant">
<contributor label="VW offered to build it">
<part label="VW offered to build it" start="3775" end="3797"/>
</contributor>

</scu>
<scu uid="18" label="When GM protested, private mediation was set up between the companies">
<contributor label="When GM protested...private mediation was set up between the two companies">
<part label="When GM protested" start="419" end="436"/>
<part label="private mediation was set up between the two companies" start="472" end="526"/>
</contributor>
</scu>

Appendix C

Pyramid Annotation for System Summary of Document Collection D324

The following provides the Pyramid annotation of the system summary by System 6 on Document Collection D324 as annotated by Annotator I. The first part contains the system summary (same as in Appendix A). Subsequently, the peerSCUs are listed, where the uid indicates the corresponding SCU from the Pyramid created from the reference summaries (cf. Appendix B). If the system summary contains the SCU, then the peerSCU contains a contributor and the start and end of the contributors are character offsets into the text in the <text></text> element.

<text>

He was drawn into the fray by Mr Ferdinand Piech, VW chairman, after Opel said it would agree to written exchanges with the VW chief if he publicly retracted statements that the US group was conducting a 'war' with the intention of 'defaming' Volkswagen.

It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr Jose Ignacio Lopez de Arriortua, stole industrial secrets from the US group and took them with him when he joined VW last year.

'It would be good if Mr Piech (VW chief executive) and Mr (Klaus) Liesen (chairman of the VW supervisory board) would send a letter to VW employees saying that, without any danger to their job security or their relationship with their supervisors, they should say what they know of this case'.

Possible charges against Mr Lopez, his colleagues and the company include wire or

mail fraud, which carry possible penalties of imprisonment or fines.
VW's weekend statement reiterated that the company remained convinced that no GM data had come into its possession or was used to its benefit.
Mr Lopez still maintained his innocence and had wanted a full hearing on the issue, his lawyers said.
He has admitted having ordered the destruction of 'possibly sensitive' GM material at VW.
Mr Lopez, just back from holiday, is expected to be questioned soon by public prosecutors on suspicions of his involvement in industrial espionage.
</text> <peerscu uid="25" label="(7) As of October 1994, nothing had been resolved">
</peerscu>
<peerscu uid="5" label="(7) Lopez left GM">
<contributor label="former GM...Mr Jose Ignacio Lopez de Arriortua">
<part label="former GM" start="411" end="420"/>
<part label="Mr Jose Ignacio Lopez de Arriortua" start="431" end="465"/>
</contributor>
</peerscu>
<peerscu uid="19" label="(7) lopez took documents from GM to VW">
<contributor label="stole industrial secrets from the US group and took them with him">
<part label="stole industrial secrets from the US group and took them with him" start="467" end="532"/>
</contributor>
</peerscu>
<peerscu uid="7" label="(7) there was an industrial espionage case involving GM and VW">
<contributor label="Mr Lopez, just back from holiday, is expected to be questioned soon by public prosecutors on suspicions of his involvement in industrial espionage">
<part label="Mr Lopez, just back from holiday, is expected to be questioned soon by public prosecutors on suspicions of his involvement in industrial espionage" start="1342" end="1488"/>
</contributor>
</peerscu>
<peerscu uid="4" label="(7) VW hired Jose Ignacio Lopez">
<contributor label="when he joined VW last year">

<part label="when he joined VW last year" start="533" end="560"/>
</contributor>
</peerscu>
<peerscu uid="15" label="(6) Ferdinand Piech is VW chairman">
<contributor label="Mr Ferdinand Piech, VW chairman">
<part label="Mr Ferdinand Piech, VW chairman" start="30" end="61"/>
</contributor>
<contributor label="Mr Piech (VW chief executive)">
<part label="Mr Piech (VW chief executive)" start="583" end="612"/>
</contributor>
</peerscu>
<peerscu uid="6" label="(6) seven other GM executives left with Lopez">
</peerscu>
<peerscu uid="29" label="(6) The FBI is probing possible mail and wire fraud">
</peerscu>
<peerscu uid="20" label="(6) the missing documents described plans to build a new model car">
</peerscu>
<peerscu uid="21" label="(5) documents included secret plans for a new factory">
</peerscu>
<peerscu uid="14" label="(5) Ferdinand Piech recruited the General Motors/Opel executive, Jose Lopez de Arriortua">
</peerscu>
<peerscu uid="30" label="(5) Lopez paid a fine instead of facing perjury charges">
</peerscu>
<peerscu uid="26" label="(5) Lopez was accused of perjury">
</peerscu>
<peerscu uid="11" label="(5) Lopez was GMs' procurement chief">
<contributor label="director">
<part label="director" start="421" end="429"/>
</contributor>
</peerscu>
<peerscu uid="12" label="(4) Adam Opel is subsidiary of GM">
</peerscu>

<peerscu uid="23" label="(4) documents were found where former GM employee were staying">
</peerscu>
<peerscu uid="24" label="(4) German officials began investigating VW for theft">
</peerscu>
<peerscu uid="43" label="(4) Lopez left Opel On March 16, 1993">
</peerscu>
<peerscu uid="10" label="(4) Lopez was Basque-born">
</peerscu>
<peerscu uid="52" label="(4) Lopez was disappointed by GM's decision not to build an automobile plant in his own Basque country">
</peerscu>
<peerscu uid="8" label="(4) VW and Lopez also were accused on conducting an illegal recruiting campaign">
</peerscu>
<peerscu uid="287" label="(3) Gutierrez and Piazza were the former GM associates who were found with the plans">
</peerscu>
<peerscu uid="40" label="(3) A regional court in Frankfurt issued an injunction preventing VW from recruiting more GM staff">
</peerscu>
<peerscu uid="41" label="(3) All charges of anti-competitive staff poaching were later dismissed">
</peerscu>
<peerscu uid="44" label="(3) Ferdinand Piech took over an ailing VW company that was losing money">
</peerscu>
<peerscu uid="35" label="(3) The U.S. Justice Department's interest in industrial espionage had been piqued">
</peerscu>
<peerscu uid="58" label="(2) Gunter Rexrodt is the German economics minister">
</peerscu>
<peerscu uid="27" label="(2) Documents included details of Opel's entire European component supplier network and key contact data">
</peerscu>

<peerscu uid="90" label="(2) Eurothere are fears of destabilization of relations between Germany and America">
</peerscu>
<peerscu uid="82" label="(2) German newspapers such as Der Speigel made public allegations of spying against Lopez">
</peerscu>
<peerscu uid="65" label="(2) GM charged that during his last months at GM, Lopez stole GM plans">
</peerscu>
<peerscu uid="57" label="(2) Gunter Rexrodt, had tried to be a peacemaker is this controversy">
</peerscu>
<peerscu uid="69" label="(2) investigation is bogged down in political and legal translatlantic issues">
</peerscu>
<peerscu uid="28" label="(2) Lopez helped turn around Opel">
</peerscu>
<peerscu uid="66" label="(2) Lopez paid fine in May, 1994">
</peerscu>
<peerscu uid="16" label="(2) Lopez was hired as VW production director">
</peerscu>
<peerscu uid="87" label="(2) Piech got nationalistic in his accusations">
<contributor label="the US group was conducting a 'war' with the intention of 'defaming' Volkswagen">
<part label="the US group was conducting a 'war' with the intention of 'defaming' Volkswagen" start="174" end="253"/>
</contributor>
</peerscu>
<peerscu uid="53" label="(2) Still Later German police raided VW headquarters">
</peerscu>
<peerscu uid="285" label="(2) The factory in stolen plans was high-speed">
</peerscu>
<peerscu uid="286" label="(2) The factory in stolen plans was low-cost">
</peerscu>
<peerscu uid="77" label="(2) The German prosecutor was Dorthea Holland">

</peerscu>
<peerscu uid="22" label="(2) Then state prosecution officials discovered four boxes of papers">
</peerscu>
<peerscu uid="32" label="(2) There were investigations and counter charges on both sides of the ocean">
</peerscu>
<peerscu uid="70" label="(2) VW counter-charged that GM had planted GM documents and data in VW sites and computers">
</peerscu>
<peerscu uid="59" label="(1) in September 1993, Rexrodt withdrew as peacemaker">
</peerscu>
<peerscu uid="50" label="(1) VW failed to convince GM that its plans were not plagiarized">
</peerscu>
<peerscu uid="295" label="(1) A court case was also brought against leading news magazine, Der Spiegel">
</peerscu>
<peerscu uid="36" label="(1) A US probe of the investigation started at the instigation of the Commerce Department">
</peerscu>
<peerscu uid="60" label="(1) A VW employee said she had punched Opel data into the VW computer">
</peerscu>
<peerscu uid="51" label="(1) As early as December, 1992, Lopez was in touch with Piech">
</peerscu>
<peerscu uid="67" label="(1) At a VW meeting in August, 1993, Lopez contradicted his earlier public claim that he never took any secret documents">
</peerscu>
<peerscu uid="78" label="(1) Because of leaks, a gag was placed on Holland's office">
</peerscu>
<peerscu uid="394" label="(1) defensive allegations were made by VW Chairman Ferdinand Piech">
</peerscu>

<peerscu uid="296" label="(1) Der Spiegal is a leading news magazine">
</peerscu>
<peerscu uid="83" label="(1) Der Spiegel later presented evidence in state court in a bid not to be stopped from reporting">
</peerscu>
<peerscu uid="42" label="(1) Documents were shredded">
</peerscu>
<peerscu uid="76" label="(1) Dorthea Holland, was searching through an estimated 2 million computer printout sheets">
</peerscu>
<peerscu uid="34" label="(1) FBI investigation was also stalled">
</peerscu>
<peerscu uid="56" label="(1) General Motors Corporation and Volkswagen were warring in 1993 and 1994">
</peerscu>
<peerscu uid="392" label="(1) Gerhardt Schroeder is from Lower Saxony">
</peerscu>
<peerscu uid="63" label="(1) Gerhardt Schroeder supported VW">
</peerscu>
<peerscu uid="288" label="(1) German politicians called the case biased">
</peerscu>
<peerscu uid="289" label="(1) Germans wanted the case dropped">
</peerscu>
<peerscu uid="298" label="(1) Germany distanced itself from Piech's accusations">
</peerscu>
<peerscu uid="94" label="(1) Germany is concerned about the effect of the court investigations on domestic economic and political affairs">
</peerscu>
<peerscu uid="17" label="(1) GM employees leaving for VW included Lopez's close friend, Jorge Manuel Gutierrez">
</peerscu>
<peerscu uid="55" label="(1) GM settled for only some employess to be banned from working for VW">
</peerscu>

<peerscu uid="54" label="(1) GM wanted all former employees banned from working for VW for 12 months">
</peerscu>
<peerscu uid="62" label="(1) Gunter Rexrodt was concerned of damage to US-German relations">
</peerscu>
<peerscu uid="85" label="(1) investigations in both countries were followed by civil and criminal court cases">
</peerscu>
<peerscu uid="292" label="(1) investigations were launched against Lopez's 22-year old daughter">
</peerscu>
<peerscu uid="294" label="(1) investigations were launched against other GM colleagues now at VW">
</peerscu>
<peerscu uid="38" label="(1) Japanese car import quotas to Europe expire in two years">
</peerscu>
<peerscu uid="80" label="(1) Lopez faced further charges">
<contributor label="Possible charges against Mr Lopez, his colleagues and the company include wire or mail fraud, which carry possible penalties of imprisonment or fines">
<part label="Possible charges against Mr Lopez, his colleagues and the company include wire or mail fraud, which carry possible penalties of imprisonment or fines" start="856" end="1005"/>
</contributor>
</peerscu>
<peerscu uid="86" label="(1) Lopez left GM under a cloud of confusion">
</peerscu>
<peerscu uid="68" label="(1) Lopez said that papers from his former offices were destroyed in order to keep them from being circulated within VW">
<contributor label="He has admitted having ordered the destruction of 'possibly sensitive' GM material at VW">
<part label="He has admitted having ordered the destruction of 'possibly sensitive' GM material at VW" start="1252" end="1340"/>

</contributor>
</peerscu>
<peerscu uid="48" label="(1) Lopez tried to recruit others">
</peerscu>
<peerscu uid="39" label="(1) Lopez was a renowned cost-cutter">
</peerscu>
<peerscu uid="9" label="(1) Lopez was eccentric and visionary">
</peerscu>
<peerscu uid="79" label="(1) Lopez was found innocent during his first trial in Germany">
</peerscu>
<peerscu uid="293" label="(1) Lopez's daughter Begounia is 22 years old">
</peerscu>
<peerscu uid="393" label="(1) Lower Saxony is VW's largest shareholder">
</peerscu>
<peerscu uid="47" label="(1) people leaving GM with Lopez included Jose Gutierrez, Jorge Alvarez Aguirre, and Rosario Piazza">
</peerscu>
<peerscu uid="45" label="(1) Piech became VW chairman in January 1993">
</peerscu>
<peerscu uid="49" label="(1) Piech was a ruthless restructurer">
</peerscu>
<peerscu uid="75" label="(1) Piech was impressed with Lopez">
</peerscu>
<peerscu uid="72" label="(1) Piech's clumsy, halfhearted conciliation efforts failed">
</peerscu>
<peerscu uid="297" label="(1) Piecs's accusations angered GM">
</peerscu>
<peerscu uid="37" label="(1) President Clinton decided that industrial espionage was a threat to America's well being">
</peerscu>
<peerscu uid="92" label="(1) Ron Brown, suggested that relations between the U.S. and Germany would be damaged">
</peerscu>

<peerscu uid="291" label="(1) Soon after Lopez's arrival, GM planned a car similar to a planned GM model">
</peerscu>
<peerscu uid="290" label="(1) Soon after Lopez's arrival, VW announced a new plant to be built in Basque country">
</peerscu>
<peerscu uid="84" label="(1) The case reached the highest levels in both countries">
</peerscu>
<peerscu uid="91" label="(1) U.S. Commerce Secretary is Ron Brown">
</peerscu>
<peerscu uid="93" label="(1) US wants German investigatorsto immediately deliver long promised data and assistance in the GM/VW case">
<contributor label="It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that">
<part label="It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that" start="255" end="410"/>
</contributor>
</peerscu>
<peerscu uid="73" label="(1) VW failed to get court injunctions preventing Der Spiegel magazine from publishing GM's allegations">
</peerscu>
<peerscu uid="61" label="(1) VW offered to build the Basque plant">
</peerscu>
<peerscu uid="18" label="(1) When GM protested, private mediation was set up between the companies">
</peerscu>
<peerscu uid="0" label="All non-matching SCUs go here">
<contributor label="'It would be good if...and Mr (Klaus) Liesen (chairman of the VW supervisory board) would send a letter to VW employees saying that, without any danger to their job security or their relationship with their supervisors, they should say what they know of this case'">
<part label="'It would be good if" start="562" end="582"/>
<part label="and Mr (Klaus) Liesen (chairman of the VW supervisory board) would

send a letter to VW employees saying that, without any danger to their job security or their relationship with their supervisors, they should say what they know of this case’" start="613" end="854"/>
</contributor>
<contributor label="Mr Lopez still maintained his innocence and had wanted a full hearing on the issue, his lawyers said">
<part label="Mr Lopez still maintained his innocence and had wanted a full hearing on the issue, his lawyers said" start="1150" end="1250"/>
</contributor>
<contributor label="VW’s weekend statement reiterated that the company remained convinced that no GM data had come into its possession or was used to its benefit">
<part label="VW’s weekend statement reiterated that the company remained convinced that no GM data had come into its possession or was used to its benefit" start="1007" end="1148"/>
</contributor>
<contributor label="He was drawn into the fray by...after Opel said it would agree to written exchanges with the VW chief if he publicly retracted statements that">
<part label="He was drawn into the fray by" start="0" end="29"/>
<part label="after Opel said it would agree to written exchanges with the VW chief if he publicly retracted statements that" start="63" end="173"/>
</contributor>
</peerscu>

Appendix D

Pyramid Annotation Instructions

This appendix presents the official Pyramid annotation instructions from (Nenkova et al., 2006).

Summarization Content Units (SCUs).

The goal of SCU annotation is to identify sub-sentential content units that can allow for comparison of the information in several summaries. It is well-known that when summarizing people make different choices about what information to include in their summary. The SCU annotation aims at highlighting what people agreed on. After the annotation is completed, some SCUs might appear in only one summary, but its annotation will allow a person to read a brand new summary and look for that SCU in this new summary.

An SCU consist of a label and contributors. The label is a concise English sentence that states the semantic meaning of the content unit. The contributors are snippet(s) of text coming from the summaries that show the wording used in a specific summary to express the label. It is possible for an SCU to have a single contributor, in the case when only one of the analyzed summaries expresses the label of the SCU.

The definition of content unit is somewhat fluid – it can sometimes be a single word but it is usually bigger than a clause. Any event realized by a verb or a nominalized verb (e.g, "blow up" and "bombing" in the examples below) is a candidate SCU.

The three questions that will help you identify an SCU contributor are

1. Is the information expressed by it repeated in some other summary? Note, the wording need not be the same for the expressed meaning to be the same; we are looking for the same meaning. When an information unit is expressed in two or more summaries, the amount of information overlap will serve as a main indication of which

parts of the corresponding sentences will become contributors. 2. Spans of words that indicate location or time, or otherwise provide more specific information about another SCU are also SCUs. Usually these are expressed in adjuncts such as prepositional phrases and are not an obligatory argument to any verb. Noun phrases containing premodification can also be split into more than one SCU when the premodifiers include additional information. For example, if the summaries under annotation convey that there was a bombing and the location of the bombing, then the annotator would identify two SCUs, one with the main event, and one with the additional detail information. 3. Is the difference important for the story? Occasionally there will be minor differences in wording that if put under scrutiny could be construed to have different nuances. We are not interested in the finest grained distinctions—these will be too many to describe in a reasonable way.

Overall, the annotation involves semantic judgements and it is thus difficult to list all possible syntactic constructions that can give rise to a content unit. The goal is to split the text in small semantic units that the original summary writers have put together in several sentences to form their summary. During the annotation, the context of the sentence and the entire summary can be used to interpret a specific text segment.

Example 1: The three sentences below come from four different summaries A, B, C and D.

A: In 1992 the U. N. voted sanctions against Libya for its refusal to turn over the suspects.

B: The United Nations imposed sanctions on Libya in 1992 because of their refusal to surrender the suspects.

C: The U.N. imposed international air travel sanctions on Libya to force their extradition.

D: Since 1992 Libya has been under U.N. sanctions in effect until the suspects are turned over to United States or Britain.

Among other information, all four sentences express the fact that "Libya was under U.N. sanctions" and this is the label for the SCU. The contributors are marked in brackets below (ignore SCU2 for now.)

A: In 1992 [the U. N. voted sanctions against Libya]1 [for its refusal to turn over the suspects.]2

B: [The United Nations imposed sanctions on Libya]1 in 1992 [because of their refusal to surrender the suspects.]2

C: [The U.N. imposed]1 international air travel sanctions on Libya [to force their

extradition.]2

D: Since 1992 [Libya has been under U.N. sanctions]1 [in effect until the suspects are turned over]2 to United States or Britain.

Other information, such as when the sanctions where imposed, what specific sanctions were imposed, why they were imposed etc, will form their own SCUs. Identifying a main topic event in the summaries and asking yourself such questions as above about specifics will help you formulate labels and identify the SCU contributors. The contributors of an SCU need not share identical wording. For example in the sentences above, the SCU with label "The goal behind the sanctions is to make Libya surrender the suspects" is expressed by the text coindexed with "2". Sentence B differs in wording from the rest of the sentences, but the meaning is the same as that of the other contributors, expressing the fact that Libya does not want to surrender the suspects and the other nations involved want to force their extradition. (Note that this is an example of only two SCUs that will be derived from the sentences, the full analysis will lead to identifying more SCUs and will lead to complete bracketing of the sentences.) Let's look at one more example of sentences from the different summaries that share some common information.

A. In 1998 [two Libyans indicted]1 [in 1991]2 for the Lockerbie [bombing]3 were still in Libya.

B. [Two Libyans were indicted]1 [in 1991]2 [for blowing up]3 [a Pan Am]5 [jumbo jet]4 over Lockerbie, Scotland in 1988.

C. [Two Libyans, accused]1 by the United States and Britain [of bombing]3 [a New York bound]6 [Pan Am]5 [jet]4 over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.

D. [Two Libyan suspects were indicted]1 [in 1991]2.

All share the information that (1) "Two Libyans are held responsible for a crime". The contributors are surrounded by brackets and coindexed by 1. Note that C differs in its wording from the other sentences–accused is not the same as indicted. But because the goal of the annotation is to find as much shared information as possible, and the sense of "accused" is so close to that of "indicted", the contributors will be grouped together, and the label expresses the general meaning of both accused and indicted.

The time expression prepositional phrase "in 1991" forms a separate SCU because the phrase "in 1991" can be omitted for example from sentence D without making the sentence ungrammatical or incomprehensible. There will be loss of information, and

this is why the phrase can indicate a new *content* unit! The contributors of the SCU with label "The libyans were accused in 1991" are coindexed with "2".

Now we have to proceed and find what other information is repeated. For example, what was the crime committed? The different sentences give different amount of detail. When deciding where to start from–remember that the main goal is identifying the same information! All sentences agree on the fact that "the crime in question is a bombing" – the contributors are coindexed with 3.

What was bombed? "An airplane was bombed" is another SCU with index 4. This information is expressed in two bigger noun phrases " Pan Am jumbo jet" and "a New York bound Pan Am jet" but "New York bound" and "Pan Am" can be omitted and the sentences will still be acceptable, so this information will be marked in a separate content unit.

The contributors are simply a part of the sentence–not all grammatical arguments necessary to reconstruct the label will be included in the contributor. This is ok, because the label will "bring in" any argument needed.

It is best if the SCU contributor can be a complete grammatical phrase. But this is sometimes not possible, so use your best judgment in assigning the specific token boundaries of the contributor.

Some specific annotation rules

1. Length of contributors: contributors are usually not very long (average of 6 words), since the content units express small, almost atomic units of information. Thus, whenever the annotator can imagine a part of the sentence conveying additional information about an event could be removed from the sentence, this part become a candidate contributor for a new SCU. For example, a sentence like "Dr. Alan Cox was tried and convicted in Winchester, England for the attempted murder of a patient in 1992." will give rise to several content units—"Dr. Cox was convicted", "The trial was in Winchester, England", "He was tried for the attempted murder a patient" and "Dr. Cox's conviction was in 1992". The main event, location, time and additional specification each represent a content unit that could be expressed seperately in another summary.

2. General vs. specific information: often times one of the summarizer will convey more specific information than other. For example, two different summaries can contain the sentences "Dogs are used to control soccer fans" and "In Italy, dogs are used to control soccer fans" or "Dogs are used to control soccer fans in the UK". All three sentences would contribute to an SCU expressing that "Dogs are used to control soccer

fans" (three contributors), while the two specific locations, Italy and the UK will be in seperate SCUs with one contributor each. Similarly, any significant modification of a statement is split a seperate SCU. For example, the in the sentences "Birth rates have decreased" and "Birth rates have decreased by 50%", there are two content units expressed, and the fact that the decrease was by half will be split in a seperate SCU because in contributes significant new information

3. Differences in meaning: In many of the examples above we saw that near paraphrases are grouped in the same content unit. A word of caution is needed here—for example two sentences such as "China has a forced sterilization policy" and "China has a forced contraception policy" are quite similar syntactically but are semantically different, one being rather more drastic than the other, and this each would form a seperate SCU. If the two concepts were linked in the same sentence by a summarizer, as in "China has forced contraception policy, including forced sterilzation", we would have a case of general vs. specific information and the previous rule will be applicable.

4. References to time: oftentimes different people refer to the time of an event using different wording, for example, in different summaries an event can be described as having happened "in 1993", "in 1994", or "in the early 90s". Such reference would usually be grouped together in one content unit, and a match in a new summary close in meaning will match the entire SCU.

5. Most often, the clauses of a complex sentence express express different details about an event or entity, and thus the goal is to limit each content unit to a clause. Occasionally, a compex sentence doesn't really contain two seperate pieces of information, as for example the sentence "The sentence will send a message that white collar crime doesn't pay." is syntactically complex, but there are no different details and the entire sentence can be a contributor to a single content unit. 6. Once the annotation is completed, spellcheck the SCU labels, since later during peer annotation one would want to search over the labels.

Peer Annotation

Peer annotation refers to the annotation of a new summary against an existing pyramid in order to evaluate how well content in the peer summary was chosen. The goal is to identify content units that are already expressed in the pyramid, as well as the new content units in the summary that do not appear in the pyramid at all. Content that corresponds to highly-weighted SCUs in the model pyramid are better, since they express information that many human summarizers have agreed on including.

The annotation is similar to the general SCU annotation—parts of the peer summary should be mapped to a corresponding SCU in the pyramid. The purpose is to identify in the peer summary near-paraphrases of SCUs in the model pyramid. The expressed meaning does not need to completely match the label of the SCU. In addition, the context of the entire peer summary as well as general common knowledge that an educated American might have can be used in the matching process. At the same time, avoid using information/context that you have gained by reading the pyramid and the human summaries that contributed to it. This is not really context or general common knowledge, and it will be unlikely that a new reader of the summary will have this knowledge. For example, if the summary contains anaphoric expressions that cannot be resolved within the peer summary itself, then the sentences should be annotated accordingly, without use of the possible anaphore referrent, even if this referen could be guessed if one reads the summaries that make up the pyramid. The summary is meant to be read without knowledge of the input documents/other summaries, so if the references are bad and the summary is unclear, then the annotation should reflect this and such vague snippet of text should not be matched to an SCU. In summary—resolving anaphora within the peer summary is ok, and it should be done during annotation (since a reader of the summary will have it available), but knowldge coming from the pyramid summaries should not be used. One can use general knowledge such as “France is in Europe”, which we can assume is known to any reader, and does not require reading other text on the summary topic.

For example, if a summary contains the sentence “Dogs are used to control soccer fans in Europe” and the model pyramid contains two SCUs “Dogs are used to control soccer fans” and “It is in Italy that dogs are used to control soccer fans”, then the first part of the summary sentence will match the general fact SCU, while “in Europe” can be mapped to the “in Italy” SCU. These can be considered paraphrases in the context of the sentences and by using simple common knowledge.

It is always a good idea to check the contributors of an SCU to see the annount of variation between them, this will oftentimes give you an indication that less strict mapping are possible, as in the example of annotating the year an event happened (”in 1994”, “in 1993”, in the early 90s”). The different contributors of a content unit express the same, or nearly the same information, so some difference are possible, and still a match can occur. For example, a peer sentence “The population of the world is likely to double - to more than 10bn people” can be mapped to an SCU with label “The world population will reach over 10 billion in 2050 with current growth rates”, even

though the temporal information is missing in the peer sentence. The decision to make the match can be facilitated by looking at all the contributors in the pyramid SCU and confirming that the peer expresses information that is in all of them, for example in this case this shared information can be the fact that the earth's population will reach 10 billion.

When the summary conveys new or significantly different information than the one in the model pyramid, these should be split to the appropraite contributor size and mapped to the service SCU at the bottom of the annotation panel.

You will notice that automatic summaries often repeat the same information in different sentences. Do annotate both instances. adding the corresponding contributors to the appropriate SCU. Repetition is taken into account in the final summary score.

Occasionally, the labels of two SCUs sound similar and it is dificult to understand just by reading the labels what is the difference between the two content units. In such cases, click on an SCUs to see their contributors—the contributors will give you an idea of what is the emphasis in each content unit.

The same text selection from the peer cannot be matched to two different pyramid SCUs. If a clauses expresses more than one SCU, different text spans that best represent the meaning of an SCU need to be selected.

When matching a peer contributor, try to find the most highly weighted appropriate SCU in the pyramid.

Some of the low-weight SCUs in the pyramid carry more information than the high-weighted ones, and have longer contributors. They represent less important information and you can be more liberal when matching peer content to them—it is enough for the peer to convey part of the information. The pyramid analysis is very useful in that it allows the annotator to deicde which information is important and should be split into fine-grained content units, and which information is not so important.

Content units from information that does not match a pyramid SCU: very often the peer summary contains information that is not covered by any of the pyramid SCUs. Such inforamtion should be split into content units (without assigning a lable). It is important that the "leftovers" are carefully split into segments that conform to the definition of a content unit. Again, complex sentences in which the different cluases convey different details, will be split into two seperate SCUs, as for example in "The unprecedented cold wave, which took the lives of 30 people in Eastern Europe, is finally ending" will be split in two SCUs, one about the number of victims, and another about the end of the cold wave. Complex sentences that do not convey more than one

new details about an event or entity will remain as a single SCU contributor, as for example “Many discovered upon their retirement that their pension money no longer existed.” and “Bilking a large number of people out of millions of dollars can lead to sentences that vary from ten to twenty years.”

When evaluating summaries that need to be of specific length, the final sentence is oftentimes truncated before the actual end of the sentence. If there is enough from the truncated sentence that one can get a match with a content unit, it should be annotated as a contributor. If it expresses no clear idea that can be mapped to a content unit, it should be put as a “Non-matching contributor”. If some ideas are expressed in the truncated sentence, but they do not correspond to an exisiting, add the appropriate parts to the “Non-matching SCU”. Annotation tool The new pyramid annotation tool, DUCView, (v. 1.2) is available now. Download DUCView by clicking on this link. This is a single jar file. If your browser saves the file under a different name, just rename it DUCView.jar.

When annotating multiple peers for the same pyramid, use the consistency check script to make sure that your annotations of the same sentence in different peers do not differ. The script takes all peer annotations (.pan files) and prints out each sentence that was annotated differently, as well as the content units it was matched to in each case. Make sure that the annotation is consistent across summaries. Note that occasionally, the exact same sentence can be ligitamately annotated differently in different peer summaries, because of the use of context in the annotation process. The different contexts might warrant different annotation. But make sure that the differences were intended, rather than due to tiredness and other human factors.

Appendix E

Source Code for Relevant Portions of Developed Infrastructure

The DocumentCollectionLoader loads all the relevant documents (source documents, information need, human reference summaries, system summaries) into a single datastructure (PyramidCollection) that provides links between the different connected elements.

```
public class DUC2005DocumentCollectionLoader {
    public PyramidCollection loadCollection(String pathOriginalDocs, String
     pathPyramid, String informationNeed, final String id) throws IOException {
        Document pyr = null;
        Vector<Document> system = new Vector<Document>();

        int counter = 0;
        for (File file : new File(pathPyramid).listFiles(new FilenameFilter() {
            public boolean accept(File dir, String name) {
                return name.toLowerCase().contains(id.substring(0, id.length()-1));
            }
        })) {
            org.w3c.dom.Document doc = XmlDocument.loadXML(file.getCanonicalPath());

            Element pyramid = (Element) doc.getElementsByTagName("pyramid").item(0);
            try {
                if (pyr == null) {
                    pyr = new PyramidResultTextDocumentLoader().load(pyramid, "scu",
                     SCU.class, ""+counter++);
                }
            } catch (Exception e) {
                e.printStackTrace();  //To change body of catch statement use File |
                 Settings | File Templates.
            }

            Element peer = (Element)doc.getElementsByTagName("annotation").item(0);
            try {
```

```
                system.add(new PyramidResultTextDocumentLoader().load(peer,
                 "peerscu", PeerSCU.class, ""+counter++));
            } catch (Exception e) {
                e.printStackTrace();  //To change body of catch statement use File |
                 Settings | File Templates.
            }
        }
        DocumentCollection original = new
         OriginalCollectionLoader().loadCollection(pathOriginalDocs, id);

        InformationNeed in = new
         InformationNeedCollectionLoader(informationNeed).getInformationNeed(id);

        DocumentCollection human = new DocumentCollection();
        if (pyr != null) {
            human.addDocument(pyr);
        }
        DocumentCollection systems = new DocumentCollection();
        if (system.size() > 0) {
            systems.addDocuments(system);
        }

        PyramidCollection p = new PyramidCollection(id, original, systems, human,
         in);
        p.bind();
        return p;
    }
}
```

Once the datastructure is loaded, the preprocessing is performed. The following code sample provides an example in the form of the interaction necessary for part-of-speech tagging. The preprocess() and postprocess() methods convert the datastructure into and out of the XML format required by LT-TTT2 POS tagger. The main Processor method runs the relevant command on the XML document.

```
public class POSTagProcessor extends Processor {

    protected POSTagProcessor(String command) {
        super(command);
    }
    public POSTagProcessor() {
        super(Configuration.getConfig().getAttribute("lxtools")+"libexec/thade-postag
         -Doc␣-m␣"+Configuration.getConfig().getAttribute("lxtools")+"model/memex/");
    }

    protected org.w3c.dom.Document preprocess(Document document) {
        org.w3c.dom.Document doc = new DocumentImpl();
        Element root = doc.createElement("TEXT");
        doc.appendChild(root);

        document.bind();
        Element sElement = null;
```

```
Node sentence = null;
for (Object o : document.getChildAccessor().getDirect()) {
    Node child = (Node) o;
    Node s = child.getParentAccessor().getAllFirst(Sentence.class);
    if (s == null) {

        if (child instanceof Word) {
            if (sentence != null) {
                sentence = null;
            }
            Element word = doc.createElement("w");
            for (Map.Entry<String,String> entry : child.getAttributes().
             entrySet()) {
                word.setAttribute(entry.getKey(), entry.getValue());
            }
            word.setTextContent(((PrimaryUnit)child).getText());
            root.appendChild(word);
        } else {
            if (sentence != null) {
                sElement.appendChild(doc.createTextNode(((PrimaryUnit)child).
                 getText()));
            } else {
                root.appendChild(doc.createTextNode(((PrimaryUnit)child).
                 getText()));
            }
        }
    } else if (sentence != null && s.equals(sentence)) {
        if (child instanceof Word) {
            Element word = doc.createElement("w");
            for (Map.Entry<String,String> entry : child.getAttributes().
             entrySet()) {
                word.setAttribute(entry.getKey(), entry.getValue());
            }
            word.setTextContent(((PrimaryUnit)child).getText());
            sElement.appendChild(word);
        } else {
            sElement.appendChild(doc.createTextNode(((PrimaryUnit)child).
             getText()));
        }
    } else if (sentence != null && !s.equals(sentence)) {
        sentence = s;
        sElement = doc.createElement("s");
        root.appendChild(sElement);

        if (child instanceof Word) {
            Element word = doc.createElement("w");
            for (Map.Entry<String,String> entry : child.getAttributes().
             entrySet()) {
                word.setAttribute(entry.getKey(), entry.getValue());
            }
            word.setTextContent(((PrimaryUnit)child).getText());
            sElement.appendChild(word);
        } else {
```

```
                sElement.appendChild(doc.createTextNode(((PrimaryUnit)child).
                 getText()));
            }
        } else if (sentence == null) {
            sentence = s;
            sElement = doc.createElement("s");
            root.appendChild(sElement);
            if (child instanceof Word) {
                Element word = doc.createElement("w");
                for (Map.Entry<String,String> entry : child.getAttributes().
                 entrySet()) {
                    word.setAttribute(entry.getKey(), entry.getValue());
                }
                word.setTextContent(((PrimaryUnit)child).getText());
                sElement.appendChild(word);
            } else {
                sElement.appendChild(doc.createTextNode(((PrimaryUnit)child).
                 getText()));
            }
        } else {
            System.out.println("SHOULD THIS HAPPEN????");
        }
    }
    return doc;
}

protected void postprocess(Document document, org.w3c.dom.Document doc) {

    /*try {
        Source source = new DOMSource(doc);
        StringWriter buffer = new StringWriter();
        TransformerFactory.newInstance().newTransformer().transform(source, new
         StreamResult(buffer));
        String text = buffer.toString();
        System.out.println(text);
    } catch (TransformerException e) {
        e.printStackTrace();  //To change body of catch statement use File |
         Settings | File Templates.
    }*/

    NodeList children = doc.getElementsByTagName("w");
    int i = 0;
    for (Object o : document.getChildAccessor().getDirect()) {
        Node child = (Node) o;
        if (!(child instanceof Text)) {
            Element childElement = (Element) children.item(i);

            assert childElement != null;
            NamedNodeMap attributes = childElement.getAttributes();
            for (int a = 0; a < attributes.getLength(); a++) {
                child.setAttribute(attributes.item(a).getNodeName(), attributes.
                 item(a).getNodeValue());
            }
```

```
                i++;
            } else {

            }
        }
    }
}
```

The following class performs the clustering of the identified content units. This class works in the context of human reference summaries as well as source (newswire) documents.

```
public class HierarchicalClustering {
    private double threshold;
    private GroupMembership group;
    private SyntacticMembership syntax;
    private long start;

    public HierarchicalClustering(double threshold, GroupMembership group,
     SyntacticMembership syntax) {
        this.threshold = threshold;
        this.group = group;
        this.syntax = syntax;
    }

    public Collection<ScuCombination> cluster(DocumentCollection coll) {
        HashMap<Integer, ScuCombination> current = createScuCombindations(coll);
        double[][] similarityMatrix = initialize(current);
        boolean foundMatch;
        do {
            foundMatch = false;
            double max = -1;
            int column = 0, row = 0;
            for (int i = 0; i < similarityMatrix.length; i++) {
                for (int j = 0; j < similarityMatrix[i].length; j++) {
                    if (i != j && similarityMatrix[i][j] > max) {
                        max = similarityMatrix[i][j];
                        row = i;
                        column = j;
                    }
                }
            }
            if (max > threshold) {
                foundMatch = true;
                similarityMatrix = combine(row, column, current, similarityMatrix);
            }
            if (current.size() % 10 == 0) {
                System.out.println("current.size():␣"+current.size()+"␣time␣so␣far:␣
                 "+(System.currentTimeMillis()-start)/1000+ "␣seconds");
            }
        } while (foundMatch);
        return current.values();
    }
```

```
private double[][] combine(int row, int column, HashMap<Integer, ScuCombination>
 current, double[][] similarityMatrix) {
    ScuCombination comb = current.remove(row);
    current.get(column).add(comb);
    for (int i = 0; i < similarityMatrix[row].length; i++) {
        if (current.get(i) != null) {
            double temp = computeSimilarity(current.get(column), current.get(i));
            similarityMatrix[column][i] = temp;
            similarityMatrix[i][column] = temp;
        } else {
            similarityMatrix[column][i] = -1;
            similarityMatrix[i][column] = -1;
        }
        similarityMatrix[row][i] = -1;
        similarityMatrix[i][row] = -1;
    }
    return similarityMatrix;
}

private double[][] initialize(HashMap<Integer, ScuCombination> current) {
    double[][] matrix = new double[current.size()][current.size()];
    for (int i = 0; i < matrix.length; i++) {
        for (int j = 0; j < matrix[i].length; j++) {
            matrix[i][j] = computeSimilarity(current.get(i), current.get(j));
        }
    }
    return matrix;
}

private double computeSimilarity(ScuCombination first, ScuCombination second) {
    return (first.computeSimilarity(second) + second.computeSimilarity(first)) /
     2.0;
}

private HashMap<Integer, ScuCombination>
 createScuCombindations(DocumentCollection coll) {
    HashMap<Integer, ScuCombination> combs = new HashMap<Integer,
     ScuCombination>();
    int counter = 0;
    for (Document doc : coll.getDocuments()) {
        long start = System.currentTimeMillis();
        System.out.println("starting␣doc");
        Collection<Instantiation> insts =
         syntax.getAllAvailableInstantiations(doc);
        System.out.println("time␣for␣doc:␣
         "+(System.currentTimeMillis()-start)/1000);
        for (Instantiation inst : insts) {
            ScuCombination comb = new ScuCombination(counter++, inst, syntax,
             group);
            combs.put(comb.id, comb);
        }
    }
```

```
        return combs;
    }
}
```

The following matches a Pyramid (cluster of content units) into the documents.

```
public class MatchClustering {
    private double threshold;
    private Mode mode;
    private double thresholdScuMatching;
    GroupMembership group;
    SyntacticMembership syntactic;

    public MatchClustering(double threshold, Mode mode, SyntacticMembership
     syntactic, GroupMembership group, double thresholdScuMatching) {
        this.threshold = threshold;
        this.mode = mode;
        this.syntactic = syntactic;
        this.group = group;
        this.thresholdScuMatching = thresholdScuMatching;
    }



    public static enum Mode {
        ANY_ABOVE_THRESHOLD, HIGHEST_ABOVE_THRESHOLD
    }

    /**
     * annotations are added to coll
     *
     * @param coll
     * @param scuCombs
     * @param name
     */
    public void annotateConcepts(DocumentCollection coll, Collection<ScuCombination>
     scuCombs, String name) {
        Vector<Template> templates = TemplateLoader.getTemplates();
        for (Document doc : coll.getDocuments()) {
            for (Sentence sentence :
             doc.getParentAccessor().getAll(Sentence.class).getSorted()) {
                for (Template template : templates) {
                    HashSet<Instantiation> insts = template.instantiate(sentence);
                    for (Instantiation inst : insts) {
                        ScuCombination scuComb = new ScuCombination(-1, inst,
                         syntactic, group);
                        double max = 0;
                        ScuCombination maxComb = null;
                        for (ScuCombination cluster : scuCombs) {
                            double temp = (scuComb.computeSimilarity(cluster) +
                             cluster.computeSimilarity(scuComb)) / 2;
                            if (temp > max) {
                                max = temp;
                                maxComb = cluster;
```

```
                        }
                    }
                    if (max > threshold) {
                        PeerSCU peer = new PeerSCU(doc, maxComb.id + "");
                        peer.setTypeSCU(name);
                        for (Map.Entry<Integer, Node> entry :
                         inst.getMapping().entrySet()) {
                            for (Word word :
                             entry.getValue().getChildAccessor().getAll(Word.class).
                             getSorted()) {
                                peer.addChild(word);
                            }
                        }
                    }
                }
            }
        }
    }

    public String annotateSCUs(DocumentCollection coll, Collection<SCU> scus, String
     type) {
        String evalName = type + "_MODE_" + mode.toString() + "_SYNTACTIC_" +
         syntactic.getClass().getName() + "_GROUP_" + group.getClass().getName();

        long start = System.currentTimeMillis();

        //obtain all the instantiations for the templates in each doc and convert to
         concept
        HashMap<Document, HashMap<Sentence, HashSet<ScuCombination>>> instsDoc = new
         HashMap<Document, HashMap<Sentence, HashSet<ScuCombination>>>();
        for (Document doc : coll.getDocuments()) {
            instsDoc.put(doc, new HashMap<Sentence, HashSet<ScuCombination>>());
            for (Sentence s :
             doc.getParentAccessor().getAll(Sentence.class).getSorted()) {
                System.out.println("current time: "+
                 (System.currentTimeMillis()-start)/1000);
                instsDoc.get(doc).put(s, new HashSet<ScuCombination>());
                for (Instantiation i : getInsts(s)) {
                    instsDoc.get(doc).get(s).add(new ScuCombination(1, i, syntactic,
                     group));
                }
            }
        }

        //obtain all the instantiations for the contributors in the human annotation
         and convert to concept
        HashMap<Contributor, HashSet<ScuCombination>> instsScu = new
         HashMap<Contributor, HashSet<ScuCombination>>();
        for (SCU scu : scus) {
            for (Contributor con :
             scu.getChildAccessor().getAll(Contributor.class).getSorted()) {
```

```
            System.out.println("contributor - current time: "+
             (System.currentTimeMillis()-start)/1000);
            HashSet<Instantiation> insts = getInsts(con);
            instsScu.put(con, new HashSet<ScuCombination>());
            for (Instantiation i : insts) {
                instsScu.get(con).add(new ScuCombination(2, i, syntactic,
                 group));
            }
        }
    }
    System.out.println("total time instantiations: "+
     (System.currentTimeMillis()-start)/1000);

    //determine the matching instantiations between the document and any
     particular contributor and determine whether
    //there is sufficient overlap
    for (Map.Entry<Document, HashMap<Sentence, HashSet<ScuCombination>>>
     entryDoc : instsDoc.entrySet()) {
        for (Map.Entry<Sentence, HashSet<ScuCombination>> entrySentence :
         entryDoc.getValue().entrySet()) {
            for (Map.Entry<Contributor, HashSet<ScuCombination>> contr :
             instsScu.entrySet()) {
                double counter = 0;
                HashSet<Word> words = new HashSet<Word>();

                for (ScuCombination combContr : contr.getValue()) {
                    double max = 0;
                    HashSet<Word> temp = new HashSet<Word>();
                    for (ScuCombination combDoc : entrySentence.getValue()) {
                        double sim = combContr.computeSimilarity(combDoc);
                        if (mode.equals(Mode.ANY_ABOVE_THRESHOLD) && sim >
                         threshold) {
                            counter++;
                            for (Instantiation i : combDoc.getOriginals()) {
                                for (Map.Entry<Integer, Node> map :
                                 i.getMapping().entrySet()) {
                                    words.addAll(map.getValue().getChildAccessor().
                                     getAll(Word.class));
                                }
                            }
                        } else if (mode.equals(Mode.HIGHEST_ABOVE_THRESHOLD)) {
                            if (sim > threshold & sim > max) {
                                max = sim;
                                temp = new HashSet<Word>();
                                for (Instantiation i : combDoc.getOriginals()) {
                                    for (Map.Entry<Integer, Node> map :
                                     i.getMapping().entrySet()) {
                                        temp.addAll(map.getValue().getChildAccessor().
                                         getAll(Word.class));
                                    }
                                }
                            }
                        }
```

```
                        }
                        if (max > 0) {
                            counter++;
                            words.addAll(temp);
                        }
                    }
                    if (counter / contr.getValue().size() > thresholdScuMatching) {
                        PeerSCU peer = new PeerSCU(entryDoc.getKey(),
                         ((SCU)contr.getKey().getParentAccessor().getAllFirst(SCU.class)).
                         getSCUID());
                        peer.setTypeSCU(evalName);
                        for (Word word : words) {
                            peer.addChild(word);
                        }
                    }
                }
            }
        }
        return evalName;
    }

    private HashSet<Instantiation> getInsts(Node node) {
        HashSet<Instantiation> insts = new HashSet<Instantiation>();
        for (Template template : TemplateLoader.getTemplates()) {
            insts.addAll(template.instantiate(node));
        }
        return insts;
    }
}
```

While the above provided excerpts from the source code for the Pyramid creation and matching, the following provides parts of the source code for the sentence ordering chapter. The first class uses the numerous possible measures for assessing the sentence ordering and annotates the documents with the results of the computation.

```
public class ProcessCoherenceMeasures {
    /**
     * args[0] - dataset name
     * args[1] - factor for combinations of coherence measures
     * args[2] - directory containing document collections
     * args[3] - directory where to save the document collections
     *
     * @param args
     * @throws IOException
     */
    public static void main(String[] args) throws Exception, ClassNotFoundException {
        String pathVO = Configuration.getConfig().getAttribute("verbocean");
        String pathWN = Configuration.getConfig().getAttribute("wordnet");

        VerbOcean vo = new VerbOcean("VerbOcean", pathVO);
        VerbOcean2 vo2 = new VerbOcean2("VerbOcean2", pathVO);
```

```
        VerbOceanNgCutoff vo3 = new VerbOceanNgCutoff(1, new HeadSimilarity(),
         "VerbOceanNgCutoff", pathVO);
        VerbOceanNgCutoff vo4 = new VerbOceanNgCutoff(1, new
         WordNetSimilarity(pathWN), "VerbOceanNgCutoff", pathVO);
        VerbOceanNgCutoff vo5 = new VerbOceanNgCutoff(2, new HeadSimilarity(),
         "VerbOceanNgCutoff", pathVO);
        VerbOceanNgCutoff vo6 = new VerbOceanNgCutoff(2, new
         WordNetSimilarity(pathWN), "VerbOceanNgCutoff", pathVO);
        VerbOceanNgCutoff vo7 = new VerbOceanNgCutoff(3, new HeadSimilarity(),
         "VerbOceanNgCutoff", pathVO);
        VerbOceanNgCutoff vo8 = new VerbOceanNgCutoff(3, new
         WordNetSimilarity(pathWN), "VerbOceanNgCutoff", pathVO);

        NounGroupCoherence head = new NounGroupCoherence("NG_Head", new
         HeadSimilarity(), 1);
        NounGroupCoherence wn = new NounGroupCoherence("NG_WN", new
         WordNetSimilarity(pathWN), 1);
        SentenceSimilarityCoherence sim = new SentenceSimilarityCoherence();
TfidfSentenceSimilarityCoherence(/*"H:\\processed081107\\processed
 \\SavedCollections",*/ /*new String[] {"C:\\Dokumente und
 Einstellungen\\Thade\\Desktop\\IdeaProjects(Linux)\\data\\original"}*/);

        ComplexNounGroupCoherence complex = new
         ComplexNounGroupCoherence("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.5);
        ComplexNounGroupCoherence complex1 = new
         ComplexNounGroupCoherence("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.25);
        ComplexNounGroupCoherence complex2 = new
         ComplexNounGroupCoherence("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.75);
        ComplexNounGroupCoherence complex3 = new
         ComplexNounGroupCoherence("ComplexNG", new WordNetSimilarity(pathWN), 1,
         1.0);
        ComplexNounGroupCoherence2 complex20 = new
         ComplexNounGroupCoherence2("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.5);
        ComplexNounGroupCoherence2 complex21 = new
         ComplexNounGroupCoherence2("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.25);
        ComplexNounGroupCoherence2 complex22 = new
         ComplexNounGroupCoherence2("ComplexNG", new WordNetSimilarity(pathWN), 1,
         0.75);
        ComplexNounGroupCoherence2 complex23 = new
         ComplexNounGroupCoherence2("ComplexNG", new WordNetSimilarity(pathWN), 1,
         1.0);
        ComplexNounGroupCoherence complexS = new
         ComplexNounGroupCoherenceSentence("ComplexNG", new
         WordNetSimilarity(pathWN), 1, 0.5);
        ComplexNounGroupCoherence complexS1 = new
         ComplexNounGroupCoherenceSentence("ComplexNG", new
         WordNetSimilarity(pathWN), 1, 0.25);
```

```
        ComplexNounGroupCoherence complexS2 = new
         ComplexNounGroupCoherenceSentence("ComplexNG", new
         WordNetSimilarity(pathWN), 1, 0.75);
        ComplexNounGroupCoherence complexS3 = new
         ComplexNounGroupCoherenceSentence("ComplexNG", new
         WordNetSimilarity(pathWN), 1, 1.0);


        GenericCoherence g13 = new GenericCoherence(new
         MinimumAverageVectorSimilarity(new SurfaceWordUnit(), new
         NGHeadSimilarityUnit()));
        GenericCoherence g14 = new GenericCoherence(new
         MinimumMaximumVectorSimilarity(new SurfaceWordUnit(), new
         NGHeadSimilarityUnit()));
        GenericCoherence g15 = new GenericCoherence(new
         MinimumMinimumVectorSimilarity(new SurfaceWordUnit(), new
         NGHeadSimilarityUnit()));


        runCoherences(args[2], Integer.parseInt(args[1]), args[0], args[3],
         complex20, complex21, complex22, complex23, complexS, complexS1, complexS2,
         complexS3, g13, g14, g15);
    }

    private static void createCombinations(int threshold, DocumentCollection coll,
     Coherence... coherences) {
        for (Document doc : coll.getDocuments()) {
            for (Coherence entry : coherences) {
                for (Coherence e : coherences) {
                    if (!e.equals(entry)) {
                        for (int i = 1; i <= threshold; i++) {
                            double valEntry =
                             Double.parseDouble(doc.getAttribute(entry.getName()));
                            double valE =
                             Double.parseDouble(doc.getAttribute(e.getName()));
                            doc.setAttribute(entry.getName() + "_plus_" + i +
                             "_times_" + e.getName(), "" + (valEntry + i * valE));
                            doc.setAttribute(e.getName() + "_plus_" + i + "_times_"
                             + entry.getName(), "" + (valE + i * valEntry));
                        }
                    }
                }
            }
        }
    }

    private static void runCoherences(String dir, int thresholdCombinations, final
     String filename, String pathSave, final Coherence... coherences) throws
     IOException, InterruptedException {
        QueueFactory.setNumberThreads(1);
        QueueFactory.initialiseQueue();
        for (File file : new File(dir).listFiles(new FilenameFilter() {
            public boolean accept(File dir, String name) {
```

```
                return name.startsWith(filename);
            }
        })) {
            if (new File(pathSave+"/"+file.getName()).exists()) {
                continue;
            }
            System.out.println("loading␣stored␣file␣" + file.getCanonicalPath());
            DocumentCollection coll =
             SaveFactory.newInstance().getSaver().loadDocumentCollection(file.
             getCanonicalPath());
            int counter = 0;
            if (coll != null && coll.getDocuments() != null) {
                for (final Document doc : coll.getDocuments()) {
                    final int c = counter++;
                    QueueFactory.getQueue().put(new QueueFactory.Entry() {
                        public void run() {
                            long start = System.currentTimeMillis();
                            if (doc.getUID().endsWith(".perm-1") ||
                             doc.getUID().endsWith(".perm-1-p") ||
                             doc.getUID().endsWith("-perm.1")) {
                                doc.setAttribute("coherenceOriginal", "1");
                            } else {
                                doc.setAttribute("coherenceOriginal", "0");
                            }
                            for (final Coherence coherence : coherences) {

                                double score = coherence.getScore(doc);
                                System.out.println(c + "␣" + coherence.getName() + "␣
                                 " + score);
                                doc.setAttribute(coherence.getName(), "" + score);

                            }
                            System.out.println("time␣for␣document:␣" +
                             (System.currentTimeMillis() - start));
                        }
                    });
                }
                QueueFactory.waitUntilQueueEmpty();
                createCombinations(thresholdCombinations, coll, coherences);
            }
            QueueFactory.waitUntilQueueEmpty();
            SaveFactory.newInstance().getSaver().save(coll, pathSave + "/" +
             file.getName());
            System.exit(100);
        }
        QueueFactory.finish();
    }
}
```

Then the SVM ranking creates the file for the learning. Then SVM^{light} can be used to learn a model and subsequently to assess the relative quality of sentence orderings.

```
public class SVMRankingQID extends SVMRanking {
```

```
private HashMap<String, Integer> uidToQid = new HashMap<String, Integer>();
private int counter = 0;
private String attributeName;
private Vector<String> features;
//private int maxTrainExamples = -1;

/**
 * qid is determined by the uid of the document
 * @param doc
 * @param input1
 * @param features
 * @param importance
 * @param qid - ignored only present for compatibility
 * @param cutoff
 * @param useZeros
 * @return
 */
protected PositionTreeSet preprocess(Document doc, StringBuffer input1,
 Vector<String> features, String importance, int qid, double cutoff, boolean
 useZeros) {
    String uid = doc.getUID();
    uid = uid.substring(uid.lastIndexOf("/")+1);
    uid = uid.substring(0, uid.lastIndexOf("perm"));
    synchronized (this) {
        if (uidToQid.get(uid) == null) {
            uidToQid.put(uid, counter++);
        }
    }
    Pattern p = Pattern.compile(".*perm\\D*1\\D*");
    if (p.matcher(doc.getUID()).matches()) {
        doc.setAttribute(importance, "1");
    } else {
        doc.setAttribute(importance, "0");
    }
    try {
    return super.preprocess(doc, input1, features, importance,
     uidToQid.get(uid), cutoff, true);
    } catch (Exception e) {
        e.printStackTrace();
    }
    return null;
}

/*public SVMRankingQID(int maxTrainExamples, Class type, String saveFile,
 boolean useCutOff, boolean useZeros, String attributeName, Vector<String>
 features) {
    this(type, saveFile, useCutOff,  useZeros, attributeName, features);
    this.maxTrainExamples = maxTrainExamples;
} */

public SVMRankingQID(Class type, String saveFile, boolean useCutOff, boolean
 useZeros, String attributeName, Vector<String> features) {
    super(type, saveFile, useCutOff, useZeros);
```

```
        this.attributeName = attributeName;
        this.features = features;
        super.features = features;
    }

    public Vector<String> getFeatures() {
        return features;
    }

    public String getAttributeName() {
        return attributeName;

    }
}
```